# Reference to objects : an empirically based study of task-oriented dialogues

*Citation for published version (APA):*
Cremers, A. H. M. (1996). *Reference to objects : an empirically based study of task-oriented dialogues*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR461534

*DOI:*
10.6100/IR461534

*Document status and date:*
Published: 01/01/1996

*Document Version:*
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

# Reference to objects

## an empirically based study of task-oriented dialogues

Anita Cremers

# Reference to objects

## an empirically based study of task-oriented dialogues

ISBN 90-386-0228-6

Omslagontwerp: Marjolein de Vette

# Reference to objects

## an empirically based study of task-oriented dialogues

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven
op gezag van de Rector Magnificus, Prof.dr. J.H. van Lint,
voor een commissie aangewezen door het College van Dekanen
in het openbaar te verdedigen
op vrijdag 14 juni 1996 om 16.00 uur

door

Anna Helena Maria Cremers

geboren te Eindhoven

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. D.G. Bouwhuis
prof.dr. H.C. Bunt

en de copromotor:

dr.ir. R.J Beun

Het onderzoek dat beschreven is in dit proefschrift is uitgevoerd aan het Instituut voor Perceptie Onderzoek (IPO) in Eindhoven, in het kader van het DenK-programma (Dialoogvoering en Kennisopbouw), een samenwerking tussen het werkverband Taal & Informatica van de Faculteit der Letteren van de Katholieke Universiteit Brabant in Tilburg, de faculteit Wiskunde & Informatica van de Technische Universiteit Eindhoven en het IPO. Het onderzoek is gefinancierd door het SamenwerkingsOrgaan Brabantse Universiteiten (SOBU).

# Acknowledgements

I wish to acknowledge the help of many people who contributed in some way to this thesis.

First, I would like to thank Robbert-Jan Beun for being my daily supervisor and an inexhaustible generator of ideas. In particular, I thank him for co-authoring chapter 3 of this thesis, and discussing and commenting on the rest of the chapters. I would also like to thank Don Bouwhuis and Harry Bunt for their constructive contributions during the course of the project, particularly in its final phase. My sincere thanks further go to (in alphabetical order) Kees van Deemter, Rudy van Hoe, Paul Piwek and Jacques Terken for comments on the contents of one or more chapters. I also thank Marc Venbrux and Retze Faber for writing the software for the study on keyboard dialogues and the experiment on spatial focus, respectively.

I should not forget to thank my present and former DenK colleagues, for sharing numerous professional, and almost as many recreational meetings (remember Corsendonk and Eersel?). Actually, the same thanks go to my present and former colleagues including my friends at IPO (remember the yearly IPO parties?). Paul Piwek deserves special thanks for enduring my daily presence as his room mate. And, Lisette Appelo, thank you for being an encouraging and emancipating monthly coach during the last year.

Further, I would like to thank the SamenwerkingsOrgaan Brabantse Universiteiten (Cooperation Centre Tilburg and Eindhoven Universities) for providing financial and personal support to carry out this research, and the Instituut voor Perceptie Onderzoek/ IPO (Institute for Perception Research) in Eindhoven for providing an excellent infrastructure as well as an agreeable scientific environment that gives you the feeling you are not the only one who either is or has been struggling to produce a thesis.

Although all these people and authorities contributed a lot to the end-product, it would have been much harder to finish it without the support and concern of my family, friends, and especially Hans, who simply refused to ever let his belief in my capabilities falter.

*Eindhoven, May 1996*

*(The butler comes back in.)*

*Butler:* Sir?

*Jacques:* Yes, O'Toole?

*Butler:* Which one is the claret, sir?

*Jacques:* The claret is in the decanter.

*Butler:* The wooden thing?

*Jacques:* No no ... the glass thing ... the glass decanter with the round glass stopper.

*Butler:* Oh yes, behind the door.

*Jacques:* No no ... on the sideboard.

*Butler:* The sideboard?

*Jacques:* The sideboard ... yes. Look ... you go into the salle à manger ... the *dining room*, right? - and the sideboard is on your left, by the wall, beside the master's portrait.

*Butler:* Ah! Above the mirror, sir?

*Jacques:* No! No! The mirror's on the other side. It's *opposite* the mirror.

*Butler:* But that's the *table*, sir.

*Jacques:* No ... you don't go as *far* as the table. You go into the room, right? ... on your right is the door to the orangery, straight ahead of you is the door to the library, and to your left is the sideboard.

*Butler:* Ah, yes, I see, sir ...

*Jacques:* And the claret is on top of the sideboard, to the left.

*Butler:* On the left.

*Jacques:* Yes ...

*Butler:* As one looks at it, sir?

*Jacques:* Yes.

*Butler:* I see, sir, thank you. (*he turns to go*)

In: G. Chapman, J. Cleese, T. Gilliam, E. Idle, T. Jones and M. Palin (1989) *Monty Python's Flying Circus: Just the words, Volume 2.* Fourth Series (Forty) nr.1, 31-10-74. London, Methuen, p. 247.

# Table of contents

# 1 Introduction

## 1.1 The goal of this thesis

In a situation where two humans cooperate in carrying out a certain task in which they each have different responsibilities, usually a dialogue will arise between them. If the participants in the dialogue can observe each other, this type of *symbolic* interaction may be extended by symbolic or other gestures, facial expressions and bodily postures. Further, if the task involves physical actions, and if one participant or both participants can observe and access the task domain, direct *physical* interaction with the domain will occur.

An important aspect of this type of task-oriented interaction is that materials and tools that are involved in the task are referred to. Given the available possibilities of interaction mentioned above, reference to these objects may be carried out by means of a natural language referring expression, a pointing gesture or a combination of the two.[1] Examples of utterances that contain one of these so-called *referential acts* to refer to blocks during a simple construction task are:[2]

(1) Do you notice the small red block on the right?
(2) Remove ↗.
(3) This one ↗ should be placed on top of that one ↗.

In example (1) the referential act is the utterance of 'the small red block on the right', in example (2) it is the pointing gesture and in example (3) the referential acts are the utterances of 'this one' and 'that one', both accompanied by pointing gestures.

The goal of this thesis is to investigate how participants in a task-oriented dialogue refer to objects that are located in a domain to which they have both visual and physical access. This means that the emphasis lies on *deictic references*, i.e., references that are used to refer to objects in the physical environment. These are usually the first references used by a participant to refer to these objects. In particular, the study focuses on the processes that determine: 1. which object the speaker chooses as the next target object, 2. the choice for a certain referential act to refer to this object, and 3. the influence of these choices on the identification process of the hearer.

The outcome of the studies, which were mainly empirical, is related to the particular characteristics of the communicative situation in which the process of object reference and identification takes place. The studies fit in the tradition of Grosz and Sidner's (1986) research on reference in dialogues about the collection of a toy water pump, using

[1] Other possibilities for the participant are to use a certain body movement or facial expression to refer to an object, but they seem to be too far-fetched to be considered any further.

[2] Pointing gestures are indicated by an arrow: '↗'.

a number of different communicative modalities. Two modalities are investigated in the present study: one in which the dialogue participants use spoken natural language and one in which they communicate via a computer terminal by typing in natural language. The domain that was chosen is a blocks domain, and the task is a simple construction task, since it was expected that basic referring mechanisms would apply here.

The choice for studying task-oriented dialogues in the type of communicative situation that was presented above was in the first place motivated by the fact that it seems to be the most natural, almost prototypical way of intention-based communication for two humans. The second motivation was that the present study is carried out in the framework of the DenK[3] project. In the DenK project a graphical and natural language interface is being developed that is based on considerations stemming from the same communicative situation. Extensive descriptions of the DenK system are provided in Ahn et al. (1995) and Bunt et al. (1995). Communication with such a system is basically task-oriented.

The DenK system is intended to embody a generic user interface design, demonstrated for the application domain of the use of an electronic microscope. Therefore, an attempt is made to extrapolate the results of the studies on the blocks domain to the domain of the electron microscope.

## 1.2 Object reference

How participants in a communicative situation as the one described above refer to objects is determined to a large extent by the context of the utterance, in particular whether the object has been mentioned before in the dialogue, and whether it is part of the visually shared domain. Bunt (1994) calls these dimensions of context the linguistic context (the surrounding linguistic material) and the semantic context (the underlying task domain, including objects, properties and relations relevant to the task), respectively. Other relevant context dimensions he mentions are the physical (the physical circumstances, including place and time of the interaction, and the communicative channels available), social (the type of interactive situation, including the roles of the participants), and cognitive (a.o., the participants' beliefs, intentions, states of processing and attention) contexts.

### 1.2.1 Referring expressions and deixis

Object reference as part of task-oriented dialogues in a visually shared domain can have at least three different functions. First of all, a speaker can employ object reference to instruct the partner to add a new object to the domain, and, accordingly, introduce it in the linguistic and semantic contexts (example (4)). Any object having the features mentioned in the description can be a referent.

(4) Please take *a large red block of two by four.*

[3] The acronym 'DenK' stands for 'Dialoogvoering en Kennisopbouw' in Dutch ('Dialogue Management and Knowledge Acquisition'), and literally means 'think'.

Second, object reference can be used to refer to objects that have already been introduced in the dialogue; these are usually objects that either are present in the shared domain or have been removed from it, These expressions are called *anaphora*, for example the expressions 'the red one', 'it' and 'the blue one' in (5):

(5) Please take a large red block and a small blue one. First take *the red one*, and put *it* on top of *the blue one*.

Third, object reference can be used to refer to an object that is present in the shared domain, and is not part of the linguistic context. This is the use of reference that is investigated in this thesis. In the domain under consideration this is usually *deictic reference*. When a speaker uses this type of reference the addressee should always take into account the physical environment (the semantic and physical contexts) to identify the referent. The blocks domain lends itself to deictic reference, since most of the objects are not unique and the physical environment is needed to distinguish the referent from other objects present. It is also possible to use non-deictic reference, though, for instance if names are given to objects. An example of deictic reference is given in (6), an example of possible non-deictic reference using a name is given in (7).

(6) Please take *the large red block on the right* and put it on *this one* ↗.
(7) Put a small red block on top of *Big Brother*.

A general definition of deixis has been provided, among others, by Lyons (1977):

> "By deixis is meant the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatio-temporal context created and sustained by the act of the utterance and the participation in it, typically, of one speaker, and at least one addressee."

This definition says that deictic expressions refer to some entity that is present in the non-linguistic context (physical or imaginary) of the utterance of which it is part. When uttering a deictic expression a speaker takes into account both his or her own position in space and time and that of the addressee, and relates these to the position of the referent.

Types of deixis that are usually distinguished are time deixis, person deixis, social deixis, discourse deixis and place deixis. In the case of *time deixis* the referent is related to the point of time at which the conversaton is being held. (e.g., 'yesterday', 'next week'). *Person deixis* is used by speakers to relate themselves to the person they are talking about (e.g., 'I', 'you'). *Social deixis* is concerned with the use of different addressing terms depending on the social relationship with the addressee. *Discourse deixis* implies reference to a part of the preceding or following text that the speaker is uttering (e.g., 'the last sentence'). Finally, *place deixis* is used to select a particular location or object included in the physical context. This type of deixis is the most relevant one in the visual task domain under consideration.

The way in which place deixis is used here is the *identifying* function, i.e., the speaker singles out an object in space by indicating its place. Other functions of place deixis that have been mentioned in the literature are the informing function and the acknowledging function (Levelt, 1989). The informing function is used to inform the hearer about the place of an object (e.g., 'the block is in front of me'). The acknowledging function is used to indicate a location that is not being referred to but that is rather presupposed (for instance, in the deictic motion verbs 'come' and 'go').

The identifying function can be performed by using demonstratives, possibly accompanied by a pointing gesture (e.g., 'this one ↗'), but also by providing information about the location of the object. For the latter type of reference both a relatum and a coordinate system may be needed. The relatum is the entity with respect to which the referent object can be localized. The coordinate system represents the field in which the demonstratum has to be found including the respective positions of the speaker, the addressee and the referent in this field.

By varying the relatum and the origin of the coordinate system, two possible references can be distinguished that are relevant to this study: primary and secondary deictic references. In the case of *primary deictic reference* the speaker is both the relatum and the origin of the coordinate system (e.g., 'the block in front of me'). *Secondary deictic reference* means that the speaker is the origin of the coordinate system, whereas the relatum is some other object. An example is 'the block behind the blue block', where 'the blue block' is the relatum.

### 1.2.2 Deictic gestures

Traditionally the gestures that people make during speech are subdivided into two categories: emblems and illustrators (Ekman and Friezen, 1972). *Emblems* are symbolic gestures that are independent of the accompanying speech, for example a waving hand to say good-bye or the 'OK'-sign. *Illustrators* are spontaneous, semi-conscious gestures whose meaning is dependent on the accompanying speech. Illustrators can occur as three different types: beats, iconix, and metaphorix (McNeill and Levy, 1982). Beats are abstract visual indicators that are particularly appropriate for emphasizing discourse-oriented functions, for example, a rising and coming down of the fingers to indicate a repair. Iconix and metaphorix can both be considered referential gestures, in the sense that they refer to an entity that is introduced in the language at the same time.

*Iconix* exhibit in form and manner of execution (e.g., forceful, slow) a meaning relevant to the simultaneously expressed linguistic meaning. An example is a speaker who says: 'he crawls up the pipe' while moving his hand upward simultaneously.

*Metaphorix* also exhibit a meaning relevant to the concurrent linguistic meaning, but the relation to the linguistic meaning is indirect. In form and manner of execution, metaphoric gestures depict the vehicles of metaphors. For instance, the metaphor 'choosing is weighing' (used in 'on the other hand' and in 'weigh the alternatives') can be expressed in a metaphoric gesture by alternating two cupped hands up and down, hereby symbolizing two possibilities that can be chosen.

The gestures that are considered in this study, i.e. the deictic gestures, are gestures that not only exhibit a semantic parallel with the accompanying linguistic unit, but also refer to some extra-linguistic entity together with this unit. They are much like iconix (or metaphorix) with the extra requirement that they are often obligatory. Without the accompanying gestures a deictic utterance can not be understood.

## 1.3 The communicative situation

Each communicative situation has three important aspects; namely, the participants in the dialogue, the domain of conversation and the modes of interaction that are available to the participants. These aspects can be represented in the form of a triangle of communication, which forms the basis of the DenK system (Figure 1).

Each communicative situation has at least two participants. These participants may differ in many aspects, particularly those aspects belonging to the social and cognitive contexts (see Bunt, 1994). For instance, the roles of participants may vary from expert to novice, and their intentions may vary from, for instance, just passing the time, or discussing a topic of common interest, to cooperatively carrying out a certain task. In the DenK triangle, the participants are the user and the cooperative assistant internal to the system.

The domain of conversation of a task-oriented dialogue may be a *mental* or a *physical* domain. The underlying task and the objects, properties and relations relevant to that task constitute, in Bunt's (1994) terms, the semantic context. In the case of a mental domain the entities that are discussed are only represented in the minds of one or both of the participants. In the case of a physical domain there are objects physically present in the communicative situation. If these objects can be perceived by the participants, they may also be represented in their minds. In the DenK system the interaction may concern both mental entities and physical objects. The mental entities of the cooperative assistant are represented in a type-theoretical context (see e.g. Ahn et al., 1995). The physical objects are represented graphically on the computer screen.

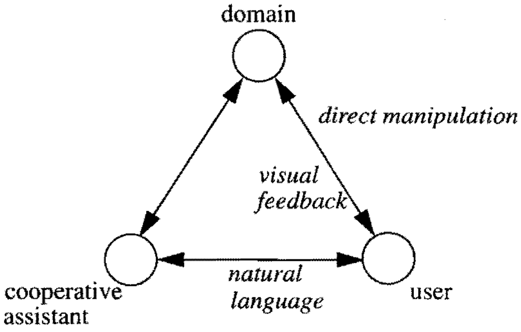


Figure 1 The DenK triangle of communication.

The possibilities of interaction available to the participants concern, on the one hand, the communication between the participants and, on the other hand, the interaction between each participant and the domain of conversation. The communication between the participants is symbolic; each dialogue participant should interpret contributions of the other participant. The main modality of communication used here is natural language (either spoken or written) that can be supplemented by non-verbal communication in face-to-face situations. In DenK the only available means of symbolic communication is natural language (English) that has to be typed in via a keyboard.

If the domain of conversation is physically present in the communicative situation, the participants may interact directly with it. They may use gestures, for instance to point at objects within the domain, and they may manipulate objects. When participants interact physically with the domain no interpreter is needed, since the domain itself is affected. In DenK the mouse can be used to point at graphical objects or to manipulate them directly.

## 1.4 The empirical studies

### 1.4.1 The situation

The empirical situation that was designed on the basis of the triangle metaphor is depicted in Figure 2 in a top-view perspective.[4] In terms of the DenK system the shared workspace on top of the table can be seen as the application domain, participant 1 as the cooperative assistant, and participant 2 as the user.

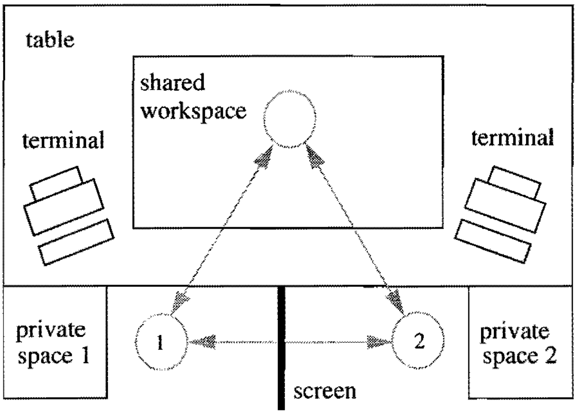


Figure 2The communicative situation (top view), derived from the triangle metaphor, that was used in two empirical studies. The terminals were used in only one of the studies.

The participants are seated side by side at the table, with a screen placed in between them, so they can not see each other's face. The screen serves as a means to prevent all non-verbal communication between the participants, except for the hand gestur-

[4] In fact, the particular set-up was very similar to the one Levinson (1992) used in his anthropological field studies.

ing. In addition, the screen hides any objects and tools that are not part of the shared workspace (yet), but that have been placed in the private workspace of either participant.

In one version of the study the participants used spoken natural language. In the second version they communicated via terminals by means of natural language they typed in on keyboards and read from the computer monitors.

The task was for participant 2 to instruct participant 1 to make some changes in a block building that was located in the shared workspace. To do this, participant 2 had an example building available in private space 2, and participant 1 had a set of separate blocks in private space 1.

### 1.4.2 A dialogue fragment

The following dialogue fragment is from the situation where the participants used spoken natural language. The original Dutch fragment has been translated into English. The '...' indicates pauses shorter than one second. If a pause was longer than one second, the exact time (in seconds) is added between brackets, for instance ...(1.7). The '--' indicates that the utterance was not completed.

```
1.  2:...(1.7) The green slide has to be removed.
2.  1:(REMOVES GREEN BLOCK)
3.  2:...Yes ...(2.3), it has to be removed completely.
4.    Eh then behind that one - the yellow one - remove it too.
      (TOUCHES YELLOW BLOCK)
5.  1:The large yellow one?
6.  2:Yes the large one, with the --
7.  1:(REMOVES A NUMBER OF BLOCKS)
8.  2:And what lies on top of it, you just remove it.
9.  1:(REMOVES A NUMBER OF BLOCKS) Yes.
10. 2:Just put it on the side, yes.
11.   ...(1.5) O dear o dear.
12.   Then you pick up a green one as big as the yellow one
      ...(1.5) was.
13. 1:Yes, (PICKS UP GREEN BLOCK) the large green one.
14. 2:Yes.
15. 1:Yes.
16. 2:Put that one across this one like this (POINTS).
17.   Here on top of the one in the back (POINTS) is that right?
18. 1:(PLACES BLOCK)
19. 2:Yes.
20.   But one more further on.
21. 1:(PLACES BLOCK) ... Yes.
```

This fragment makes clear that in this type of dialogue many deictic referring expressions and gestures occur. In order to refer to objects participants sometimes use referring expressions only, for example, the definite expression in line 1.: 'the green slide'. Combinations of referring expressions and pointing/touching gestures are also used, for example, in line 4: 'behind that one the yellow one (touches yellow block)'.

Furthermore, the fragment shows that it may take several turns before the participants agree on having identified the target object. For instance, in line 1. the instructor gives a command, it is carried out immediately by the builder in line 2. and confirmed by the instructor in line 3. Sometimes more turns are needed, for instance, if the first reference is apparently not clear enough, so that the partner has to ask for clarification, which is the case in line 5: 'The large yellow one?'.

## 1.5 The organization of this thesis

A general introduction to the empirical study of reference in spoken dialogues is provided in chapter 2. An investigation is made of the types of all the referring expressions and gestures that occurred in the spoken empirical dialogues. The resulting classification is based on the features that were used in the referring expressions. The observation that when participants refer to objects, they apparently take into account the spatial and the functional focus of attention is also reported in this chapter. This observation means that the participants only consider a certain subset, which is either spatially or functionally determined, of the objects in the domain.

In chapter 3 a different classification of referring expressions is made than the one reported in chapter 2. In this classification the types of references that are considered are limited to first references to objects that are physically present in the domain. These references are classified according to two feature types: absolute and relative features. In this chapter, the idea is introduced that participants try to use a minimal amount of effort in choosing the next object to refer to, formulating the referring expression and identifying the target object. It is demonstrated that the choice of feature types and the presence of spatial or functional focus areas play an important role in minimizing this effort.

Chapter 4 elaborates on the ideas presented in chapter 3, using the same empirical material. In this chapter the cooperative process of object reference is analyzed at utterance level. At this level the number of turns as well as the types of utterances that are used to reach mutual acceptance of the referential act and of the identification of the target object are relevant. In particular, it is demonstrated how these factors relate to the choice of features and to the focus of attention, and how they contribute to the minimization of effort.

In chapter 5, a second empirical study is introduced concerning keyboard dialogues. The first references that were used in these dialogues are again analyzed in terms of minimal effort and compared to the results of the study of spoken dialogues. The difference in the modality of communication appears to have a significant influence on the use of features in referring expressions and on the use of referring gestures. The difference also had a significant effect on the use of the focus of attention.

Chapter 6 presents the results of an experiment that was carried out to study the influence of the spatial focus of attention, assumed to play a role in the choice of referring expressions in the previous chapters. In this experiment subjects were presented with a series of simple, graphical, blocks domains on a computer monitor, each accompanied by a referring expression with one of three different grades of information. The experiment showed an influence of the spatial focus of attention depending on the grade of specification (ambiguous, minimally specified or redundant) of the referring expression involved.

In chapter 7, the current application domain of the DenK system is described, i.e., an implementation of the electron microscope. In particular, it is compared to the blocks domain in terms of features of objects and relations between them. Finally, the process of object reference as it is expected to take place in the electron microscope domain, based on the findings in the blocks domain, is described.

Chapter 8 provides an overview of seven multimedia systems that are compared to the DenK system in terms of their communicative situations. In particular, the ways in which object reference and, where applicable, focus of attention are treated in these systems is outlined.

Finally, in chapter 9, conclusions are drawn from the present thesis. Limitations in the present research and suggestions for future research are presented in terms of possible communicative situations and minimal effort.

# 2 Object reference in spoken dialogues

### *Abstract*

*Effective cooperation of humans with other humans and with intelligent machines requires a language of words and gestures that is accurate and efficient in making reference to objects. In the present investigation, expressions and gestures that were used by subjects to direct partners' attentions to building blocks during a collaborative construction task were analyzed and classified into four main categories. In addition, the influence of mutual knowledge of the participants (either knowledge about the dialogue or about the domain of conversation) on the referential acts used was studied. The focus of attention, both within the dialogue and within the domain, plays an important role in the use of references. In particular, the effect of focus within the domain of conversation on the use of referential acts needs to be investigated further.*

✳ This chapter is a slightly revised version of the paper that has been published as: Cremers, A.H.M. (1994) Referring in a shared workspace. In: M.D. Brouwer-Janse and T.L. Harrington (eds.) *Human-machine communication for educational systems design* (NATO ASI Series, Subseries F, Computer and Systems Design 129). Berlin, Springer Verlag, pp. 71-78.
✳✳ An abstract of a version of this paper has been published as:
Cremers, A. (1993) Referring behaviour in task dialogues. In: Sanford, A.J., Anderson, A.H., Moxey, L.M. and Gilhooly, K. (eds.) *Abstract of papers*. Glasgow, International Conference on The Psychology of Language and Communication, University of Glasgow, 31st August - 3rd September 1993.

## 2.1 Introduction

When two people discuss a task they are to perform together, they must indicate which of the available materials and/or tools each person is going to use. Ordinarily, each will use *referential acts,* verbal and non-verbal signals to single out each object in space, so that the partner will be able to locate it.

Research into the means of referring to objects or places in the extra-linguistic context has focused mainly on the linguistic part of referential acts uttered in isolation. It is reported that the decision of a speaker to use a particular expression to refer to an object largely depends on the type of coordinate system that is used, the place of the origin of this system and, possibly, the chosen *relatum*, i.e. the object or person that is chosen as a reference point with respect to which this object is located (Levelt, 1989). However, in face-to-face dialogues, these utterances are in many cases accompanied by non-verbal communicative acts, such as gestures, facial expressions and/or bodily postures. In addition, the form (the information that is included in the expression) and content (the object the expression refers to) of referential acts largely depend on the mutual knowledge of speaker and hearer; the *common ground,* as Clark and Wilkes-Gibbs (1986) call it. Mutual knowledge can either be general knowledge about the domain of discourse (e.g. properties of objects in the domain), knowledge about the actual state and history of the domain of discourse (e.g. the past or the present location of a certain object) or knowledge about the preceding utterances in the dialogue (e.g. a name that has been given to a certain object).

The goal of the present research is to provide an overview of types of referential expressions and gestures, and to investigate how a speaker's assumption about the hearer's knowledge influences the information that is included in subsequent references. In the following, the methodology of one dialogue empirical study that was carried out to investigate referential expressions relating to objects in a relatively restricted domain will first be described. Then, a classification will be given of the types and amounts of referential expressions and gestures that were used in the dialogues. Finally, the influence of the speakers' and the hearers' mutual knowledge on the type of referential act used will be outlined.

## 2.2 Methodology

An empirical study was carried out in which ten pairs of Dutch subjects participated. The empirical set-up and task were designed so as to evoke as many varied referential acts as possible. One of the participants (the instructor) was told to instruct the other (the builder) in rebuilding a block-building on a toy foundation plate in accordance with an example that was provided. The building consisted of blocks of one of four different colours, three sizes and four shapes. The partners were seated side by side at a table, but were separated by a screen. Only their hands were visible to one another, and only when placed on top of the table in the vicinity of the foundation plate. Both subjects were allowed to observe the building domain, to talk about it, and to gesticulate in it, but only the respective builders were allowed to manipulate blocks. The empirical configuration is depicted in Figure 3. The ten building sessions, the dialogues of which were similar to Grosz's task dialogues (Grosz, 1977), were recorded on videotape.

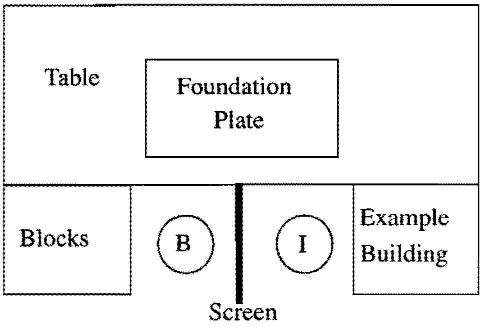


Figure 3 Empirical configuration (top view),
B = builder, I = instructor

## 2.3 Referential expressions in the dialogues

In the dialogues four main categories of reference were distinguished: reference to physical features of the object; reference to the location of the object in the domain; reference to the orientation of the object in the building domain and, finally, reference to the history which was developed in the course of actions that were carried out in the domain and the topics that were discussed in the dialogue. In the dialogues a total amount of 665 referential expressions occurred, of which 59.5% (396) actually contained information out of one or more of the four categories mentioned above. These expressions either acted as direct references to objects in the domain (62.4%, 247) or as anaphora, which means reference to objects that had already been introduced previously in the dialogue (37.6%, 149). The remaining 40.5% (269) consisted of pronominals or demonstratives; consequently these references did not contain information of any of the categories mentioned above.

### 2.3.1 Reference to physical features of an object

By far the majority of the referential expressions that were used included information about physical features of the object. Specifically, 92.2% (365) of the referential expressions included this type of information, and 77.5% (307) of the expressions consisted of physical information only. The features that were used in the dialogues were colour (e.g. 'red'[1]), size (e.g. 'large') and shape (e.g. 'square') of the blocks, which were the three distinguishing physical features in the set of blocks. There was a preference for the feature colour. This was used in 97.3% (355) of the references that contained information about physical features, whereas size and shape were mentioned in only 25.2% (92) and 17.8% (65) of the cases, respectively.

[1] Since the examples of referential expressions provided in this chapter are merely abstractions of the Dutch expressions that were used in the dialogues, they are only given in English.

## 2.3.2 Reference to the location of an object

Reference to the location of an object was mainly used by participants who did not use pointing gestures at all. This is not surprising, because both referential means seem to accomplish the same effect, one by means of language, the other through gesturing.

### 2.3.2.1 Location in general

Participants in dialogues sometimes used general expressions referring to objects in the domain, e.g. 'here' and 'there'. In 3.8% (15) of the information-containing expressions this was done, and in 1.8% (7) of the cases this reference type was used in isolation. When these references were used to refer directly (non-anaphorically) to objects in the domain, pointing gestures were always added. Most of the occurrences of this type of reference were uttered at points where the speaker shifted his or her attention to another region of the domain, for instance to give instructions about building a new part of the block-building. We shall call this kind of shift a transition of the *focus of attention.*

### 2.3.2.2 Location with respect to the participants

The location of an object was sometimes indicated by stating its relative position with respect to both participants, e.g. 'the blue one on the right'. This type of information was used in 6.1% of the references, and in 1.5% of the cases it was the only type used. These expressions were hardly ever accompanied by pointing gestures, probably because they already contained enough information to identify the intended object. Referencing by indicating the location of an object with respect to the participants was again mainly used to install a new focus of attention in the domain.

### 2.3.2.3 Location with respect to other objects

The locations of objects were also indicated as a position with respect to other objects in the domain, e.g. 'the blue block behind the yellow one'. In these expressions the position of the participants should also be taken into account. They were used in 3.5% (14) of the cases, and in 0.8% (3) of the cases this was the only type of information supplied. This type of reference was used mainly when the current block was located in the neighbourhood of the block referred to previously; in fact this happened in 71.4% (10) of the cases. In these cases the previous block was used as relatum for the current one. No transition of focus occurred, because the location was already more or less clear. This could also be reason why no pointing gestures were being used to accompany this type of expression.

### 2.3.2.4 Location with respect to the hand of a participant within the domain

The occurrence of referential expressions using the hand of a participant as a relatum was probably a consequence of the specific set-up of the task. In this set-up the participants shared the same perspective, and the only difference between each other's body positions they could actually observe was the position of the hands, which changed constantly. Instances of this type of expression occurred in 1.3% (5) of the cases, and in 0.8% (3) of the cases it was the only disambiguating information offered. No gestures accompanied these expressions.

The position of the partner's hand was used in two ways. In the first, the speaker informed the partner about where a block was located with respect to the location of his hand at the moment of the utterance, e.g. 'the blue block to your right' (to the right of your hand). In the second, the speaker told the partner in what direction his hand should move in order to reach the intended object. So, in this case, the location of the hand was presupposed, and instead of informing the partner about the location of the object, the action he or she had to carry out was supplied. This action could either be a default-action or an explicitly indicated one. In the case of a default action the partner was already moving in the right direction and only needed to be encouraged to continue doing so, e.g. 'a bit further'. In the latter case, the explicit direction had to be provided, e.g. 'go to the left'.

### 2.3.3 Reference to the orientation of an object

The different ways in which an object can be positioned in the domain all result in different object orientations. Speakers can make use of an object's orientation to distinguish it from other (identical) objects, e.g. 'the horizontal yellow block'. In 1.0% (4) of the referential expressions participants made use of this disambiguating device, but the information was always accompanied by other types of information. No accompanying gestures were used.

### 2.3.4 Reference to the history of an object

Finally, an object could be disambiguated by means of reference to earlier events in which the object had been involved, e.g. 'the red one you have just put down'. Also, when one of the participants had already talked about the object earlier in the dialogue, this could be used for disambiguation, e.g. 'the red one you just mentioned'.

In 7.1% (28) of the expressions participants made use of historical aspects, and in 2.8% (11) of the cases this was the only information they provided. References to the history of the domain and the history of the dialogue occurred equally often. Hardly any accompanying gestures were used. The ones that were used referred to objects that had been talked about before and were still located in the domain.

## 2.4 Influence of mutual knowledge on referential expressions

During the course of a dialogue and of events in the domain of conversation, knowledge is built up about these issues. Later in the dialogue and in the building process, a partner may make use of the knowledge he or she assumes the other has available. In fact, the assumption of the presence of this knowledge may to some extent determine the type of information that is being used in the current expression, although it may be fully disambiguating in its own right.

### 2.4.1 Mutual knowledge about the dialogue

Assumed knowledge about the dialogue on the part of the partner was reflected in the use of pronominal anaphora and ellipsis in the referential expressions. Anaphora can be used when the speaker assumes that the object referred to is in the focus of attention of both participants (Grosz, 1977), in which case a pronominal reference suffices for making clear which object is meant (e.g. 'take a small red block, put *it* on top of the large green one'). Ellipsis occurred when the referential expression in an utterance was omitted altogether, or when parts of the utterance were omitted. Total omissions took place when both participants had already agreed upon which object was being referred to, and only some predicate of the object was left to express, for example the destination of a particular block on which the action had already been carried out (e.g. 'take a small red block, put it on top of the large green one, (...) the green one on the right'). Partial omissions occurred when information in the current expression partially coincided with information in the preceding expression, e.g. 'place a small yellow block, put a blue one on top of it'. In this example the 'blue one' was taken to be a small block as well, although the explicit information was omitted.

### 2.4.2 Mutual knowledge about the domain: focus of attention

Knowledge about the domain of discourse and about the manipulations that are being or have been carried out in it may be reflected in referential expressions. Triggers for this type of reference that were observed in the empirical study are a participant's awareness of the existence of either a spatial or a functional focus of attention.

#### 2.4.2.1 Spatial focus

The spatial focus of the domain is the part of the domain that is being attended to most closely. For objects that are located in areas closely attended to, the speaker does not have to provide fully disambiguating information when referring. In these cases, the disambiguation should only concern the part of the domain that is being attended to. The spatial focus can be created either explicitly by verbal means, or implicitly, or by means of gestures.

**Explicit**

Speakers sometimes announced explicitly that objects that were located in a particular sub-domain should be attended to more closely, e.g. 'let's move to the upper right part'. In this way the speaker made sure that the hearer focused his or her attention on the indicated sub-domain, so that he or she could use less information in expressions referring to objects that were located there. In the dialogues, 6.1% of the referential expressions contained this type of information, and 1.5% of these consisted exclusively of this type of information.

**Implicit**

Speakers also implicitly made use of the assumed focus of attention of the partner. The speaker could have well-founded reasons for believing that the partner's focus was actually directed at this particular sub-domain, for example when he or she had just referred to an object that was located there. In this case, the speaker argued that the partner was inclined to consider the next reduced expression as a reference to an object in the neighbourhood of the one just mentioned. For example, in 'please remove these blocks (points at a green, a yellow and a blue block), the red one can remain seated there', the speaker referred to a red block without fully disambiguating it. However, it was clear to the listener that the speaker was referring to the red block in the vicinity of the three blocks that had been pointed at earlier.

**Gestures**

The partners used their hands to point at objects, to indicate their orientation in the domain, to touch them, to pick them up or to hold them. Both the instructor and the builder were allowed to point at blocks and to touch them, but the latter could in addition pick up the blocks and hold them in his or her hand. Both partners could also make use of the fact that the other was (incidentally) pointing at or touching a block. In those cases, the speaker could refer orally to that particular block by means of a minimally informative expression, like 'that one'. The instructor could also use this mechanism when the builder was picking up a block or holding it. In all of these cases the speaker could use less information than he would have needed if there had not been some involvement of a hand. In the dialogues, participants used some kind of accompanying gesture in 16.8% of the referential expressions. However, three out of ten instructors did not use any gestures at all during the dialogues. Since the participants had been instructed to act as spontaneously as possible, this can be considered a matter of preference. This lack of gesturing did not have a notable effect on the percentage, because in these dialogues the builders had to use more gestures in order to verify whether they had correctly identified a particular block. This was less necessary for builders who had a pointing instructor as a partner.

#### 2.4.2.2 Functional Focus

In addition to the assumed knowledge about spatial focus, the speakers made use of the partners' assumed knowledge about the current functional focus. Functional focus is related to the actions that have to be carried out in the domain. The concept of functional focus applies when the action that should be carried out more or less restricts the number of blocks that can reasonably be involved in the action. In that case, reduced information is possible, based on the assumed acquaintance of the partner with the pre- or post-conditions of the action.

**Preconditions**

A partner could make use of the preconditions of an action when an object had to be removed. In that case, the object referred to was most likely the one that was located at a position that was easily reachable, e.g. 'remove the little yellow one', which was taken to be the small yellow block on top of the building, although there were many other small yellow blocks available at positions that were not so easy to reach.

**Post-conditions**

Post-conditions of an action could be used when an object was to be (re-)placed. Then the object referred to was most likely the one that fitted best at the indicated location. For example, when there was an opening between two yellow blocks that could only contain a small block, the 'green block' in the utterance 'place a green block between the yellow ones' was taken to be a small one.

In total, the dialogues contain about 30 cases of the implicit spatial focusing and the functional focus mechanism together. In these cases, it can be demonstrated that the partner was actually making use of one or both of these mechanisms, because the information provided did not fully distinguish the referent object from the surrounding ones, and no pointing gestures were used. Actually, it was not always possible to distinguish between the use of implicit focusing and functional focusing, because these mechanisms coincided many times. For example, when a speaker had just been talking about a yellow block, and subsequently said that a large red block should be placed on top of a small green one, he or she probably meant the green block that was placed in the neighbourhood of the yellow block, and was suitable for having a large block placed on top of it.

## 2.5 Summary and conclusions

In this chapter, an analysis has been presented of verbal and gestural references to objects that occurred in ten task dialogues, in Dutch, and of the influence of mutual knowledge on the usage of these acts. In the dialogues, four main categories of references to objects occurred, namely reference to physical features, to the orientation or the location of objects, and to the history of objects in a domain or in the dialogue. It could be demonstrated that the participants made use of mutual knowledge about the contents of the preceding dialogue, as well as the state and history of the domain of conversation when uttering a particular referential expression.

The focus of attention appeared to be a central notion in the use and interpretation of referential expressions. At focus transitions participants tended to use more expressions referring to the location of objects in general and to locations with respect to the participants. Reference to the location of an object with respect to other objects tended to be used when the focus did not shift. Moreover, the focus of attention played an important role in the production and comprehension of references when speakers made use of mutual knowledge. The effects of the focus in dialogues on the possible use of pronouns and ellipsis are well-known (Grosz, 1977; Grosz, 1981). However, further research is needed to establish the effects of the focus in the domain of conversation on the use of references.

# 3 Object reference, minimal effort and focus of attention

**with Robbert-Jan Beun**

## *Abstract*

*In this chapter we report on an investigation into the principles underlying the choice of a particular referential expression to refer to an object located in a domain to which both participants in the dialogue have visual as well as physical access. Our approach is based on the assumption that participants try to use as little effort as possible when referring to objects. This assumption is operationalized in two factors, namely the focus of attention and a particular choice of features to be included in a referential expression.*

*We claim that both factors help in reducing the effort needed to, on the one hand, refer to an object and, on the other hand, to identify it. As a result of the focus of attention the number of potential target objects (i.e., the object the speaker intends to refer to) is reduced. The choice of a specific type of feature determines the number of objects that have to be identified in order to be able to understand the referential expression.*

*An experiment was conducted in which pairs of participants cooperatively carried out a simple block-building task, and the results provided empirical evidence that supported the aforementioned claims. Especially the focus of attention turned out to play an important role in reducing the total effort. Additionally focus acted as a strong coherence-establishing device in the studied domain.*

∗ This chapter is under revision for publication as:
Cremers, A.H.M. and Beun, R.J. (1995) Object reference in a shared domain of conversation. *IPO Manuscript 1089.* For: *Pragmatics and Cognition.*

## 3.1 Introduction

When two people discuss a task they are to perform together, they must indicate, among many other things, which of the available objects should be used. If the task is carried out in a *shared domain,* i.e., a domain to which both participants have visual as well as physical access, they can refer to these objects by means of *referential acts,* i.e. verbal referential expressions and/or nonverbal references, such as pointing or other gestures.

The primary goal of this chapter is to present some conversational principles of object reference in a shared domain of conversation. More specifically, we will be concerned with the rules underlying the choice of a particular referential act to indicate an object that has been selected by the speaker. We will call this object the *target* object. Hence, the main questions to be answered in this chapter are *how* speakers refer to a specific target object and *why* speakers opt for a specific surface structure of the referential act, given the circumstances of the utterance.

Our analysis will be based on the *principle of minimal cooperative total effort,* that takes not only the minimization of the effort to *verbalize* the expressions in a conversation into account, but also a minimization of the effort to *identify* the relevant object(s) by the hearer. Hypothetically, this minimization can be established in at least two ways. Central in our approach is the assumption that participants in a conversation establish some kind of focus space (see also, e.g., Grosz, 1977; Grosz and Sidner, 1986) that enables the speaker to use less information than actually needed when taking the complete domain of conversation into account. We also assume that by choosing a specific type of feature, a speaker can limit the number of objects that must be identified before the referential act can be understood.

Here we will focus on the part of the referential act that we call the *descriptive content.* This is the part where the speaker actually provides content information about the object to be identified, i.e., the entire referential act except the determiner and gestures. Moreover we will restrict our analysis to first references to target objects, since in those cases the identification implies explicit searching by the addressee in the shared task domain, and the descriptive content contains the maximal amount of information.

To find evidence in real discourse for the hypotheses that we formulated on the basis of the principle of minimal cooperative total effort, we conducted an empirical study where pairs of Dutch subjects had to carry out a specific task in a shared domain of conversation.

In section 3.2 we define referential acts, focusing on the descriptive content of these acts. In section 3.3 we introduce the principle of minimal cooperative total effort, which we think is the underlying mechanism for object reference. Sections 3.4 and 3.5 deal with the two important factors that follow from this principle: the focus of attention and the choice of features in the descriptive content. In these sections hypotheses are formulated about the choice of particular features in the descriptive content and the influence of the focus of attention on this choice. In section 3.6 the setup of the empirical study that was carried out is described. Section 3.7 contains the criteria that were used for determining whether or not an object was located in the current spatial focus area.

The empirical evidence for the hypotheses derived from the empirical results is discussed in section 3.8, followed by a discussion in section 3.9. Finally, in section 3.10, some conclusions are formulated.

## 3.2 Form and content of referential acts

An instance of a referential act may consist of a referential expression, possibly accompanied by a gesture. In this chapter we are only concerned with reference to single objects, so only singular expressions are considered. For our purposes, we assume a possible referential act to be constructed as in the following schema. The brackets in this schema indicate that the category is optional. However, at least one of the optional categories must be present in each rule. The star (*) indicates that the category can be used more than once. Gestures are indicated by a ↗.[1]

| | |
|---|---|
| referential act = | (referential expression) (↗) |
| referential expression = | (determiner) (descriptive content) |
| descriptive content = | (premodifier)* (head) (postmodifier)* |

Examples of referential acts are:

(1) $(het)_{det}\,((grote)_{premod}\,(rode)_{premod}\,(blok)_{head}\,(voor\ mij)_{postmod})_{descr.cont}$
'the large red block in front of me'

(2) $(een)_{det}\,((groot)_{premod}\,(blok)_{head}\,(dat\ achter\ de\ rode\ staat)_{postmod})_{descr.cont}$
'a large block lying behind the red one'

(3) $(die)_{det}\,((grote)_{premod}\,(hier)_{postmod})_{descr.cont}$ (↗)
'that large one here (↗)'

### 3.2.1 Determiners, pronouns and gestures

The first part of the referential act consists of a determiner. Determiners that can be used for making reference to single objects are indefinite articles, definite articles or demonstratives (in the examples respectively 'een' ('a'), 'het' ('the') and 'die' ('that')). In general, the use of an indefinite determiner indicates that the object referred to is being introduced in the discourse (in other words, is 'novel'). The use of a definite determiner (definite articles or demonstratives) indicates that the object is known by both participants, either because it has been mentioned before in the discourse or because it is prominent within the nonlinguistic context (in other words, is 'familiar') (Heim, 1982).

A gesture may be added to the referential expression, as is shown in example (3).

[1] Actually, in English as well as in Dutch, the form of references can be more complicated (Quirk et al., 1972; Bennis and Hoekstra, 1983). A reference may be constructed of: *(predeterminer) (determiner) (postdeterminer)* (premodifier)* (head) (postmodifier)** or *(pronoun)*. However, we will only consider the more simple form here. Moreover, although reference to objects can also be carried out by using *proper names*, such as 'De Nachtwacht' ('The Night Watch'), in this chapter we will not be concerned with these. The referential process becomes easier if objects have names assigned to them, since then there is a one-to-one relationship between the name and the object, and no alternative objects need to be considered for identification.

The type of gesture we are considered with here is the *referent-related* gesture (Knapp and Hall, 1992). According to Knapp and Hall, these gestures can be either pointing movements, drawings of the referent's shape or movement, or depictions of spatial relationships.

The above schema does not indicate that pronouns can also be used as a referential expression instead of a combination of determiner and descriptive content (e.g., 'het' ('it')). However, in this chapter we will not be concerned with pronouns, since we will concentrate on the analysis of the use of information in the *descriptive content* of the referential act. Pronouns, determiners and gestures will only be included in the analysis when necessary.

### 3.2.2 Descriptive content

The descriptive content follows the determiner in the referential act and may consist of one or more *premodifiers*, a *head,* and one or more *postmodifiers*. Premodification is carried out by means of adjectives (e.g., 'groot' ('large'), 'rood' ('red')). In contrast to English, where 'one' can be used instead of the noun, the head is usually a noun in Dutch (e.g., 'blok' ('block')). If the noun is not used in Dutch, an ellipsis takes place and the noun is omitted altogether (e.g., example (3)). Post-modification is expressed by means of a relative clause (e.g., 'dat achter de rode staat' ('that is lying behind the red one')) or a prepositional phrase (e.g., 'voor mij' ('in front of me')). We assume that predicates of the object are expressed in the pre- and post-modifiers and type information of the object in the head.

Semantically, we distinguish between *absolute* and *relative* features, both of which can be expressed in the descriptive content. Absolute features are features that can be identified without having to consider other entities; for instance, the feature 'colour' and the type of the object (e.g., 'het rode blok' ('the red block')). Relative features can be either implicit or explicit. In both cases, though, other entities have to be identified to interpret the meaning of the expression. In the implicit case, the other entities are omitted from the surface structure of the descriptive content, e.g., 'the left block', 'the large one'. In these examples the omitted entities are, respectively, the participants in the dialogue and other objects. In the explicit case, other entities are always included in the surface structure (e.g., 'the block behind the red one'). Following Levelt (1989), we will call the entity involved as a reference object the *relatum* (in our example 'the red one').

## 3.3 The principle of minimal cooperative total effort

In an actual interactive situation, the speaker may use one or more of the features described in the previous section to indicate a particular object. For instance, the speaker may refer to a specific object by saying 'the red block' or 'the block left of the yellow one'. But, which description is the most appropriate one given the cooperative situation, and are we able to formulate principles and concrete rules that predict the use of a specific description?

The first principle that comes to mind is the Gricean *maxim of quantity* (Grice, 1975)[2], which states: (1) make your contribution as informative as is required for the current purposes of the exchange, and (2) do not make your contribution more informative than is required. With respect to object reference this means that, on the one hand, speakers try not to be vague or ambiguous, since this would make it difficult for the addressee to identify the intended referent. On the other hand, they try not to be redundant, since this would cost them too much effort to utter the referring expression, and would probably confuse the addressee. It is unclear, however, how this principle can be expressed in concrete parameters that constitute a specific dialogue situation.

Other researchers have formulated the related *principle of minimal effort* (Brown, 1958; Krauss and Glucksberg, 1977; Olson, 1970). They stated that speakers try to utter a noun phrase that is as short as possible, while still allowing the hearer to select the referent. Clark and Wilkes-Gibbs extended this principle by saying that making reference to objects can be seen as a collaborative process (Clark and Wilkes-Gibbs, 1986). Their *principle of minimal cooperative effort* expresses the idea that there is a trade-off between the noun phrase that is uttered first and the possible additions or corrections to this utterance by the speaker or the partner. Hence, a speaker can decide to start by uttering an ambiguous expression, expecting the partner to make an educated guess about the intended referent or to ask for clarification if this was not possible. This results in a shared responsibility of both speaker and hearer for the establishment of the common knowledge that the expression is understood well enough for the current purposes.[3]

In this chapter, we will consider the situation in which referential expressions always contain just enough information for the partner to be able to identify the object, in line with the principles discussed above. However, in our opinion, when people are referring in a shared domain, the principle of minimal cooperative effort should be interpreted in a broader sense. The speaker and the addressee not only try to *say* as little as possible together, but they also try to *do* as little as possible. This *principle of minimal cooperative total effort*, as we call it, should not only be based on the amount of language that people use, but also on the amount of effort it takes to actually *identify* the target object.

A reduction in effort can be established in at least two ways. In the first place, the speaker can reduce the number of features in the description by trying to take as few potential target objects as possible into account. He can do this by making use of factors that are related to the focus of attention of the participants. In the second place, the speaker can try to involve as few objects as possible in the description itself. He can do this by making use of absolute features that require the identification of only one object. In the next section, implications of the focus of attention on referential behaviour will be discussed. In section 3.5 we will describe the possibilities for choosing features in the

[2] Since we consider the target object as being relevant with respect to dialogue situation and the participants as being cooperative, we will not include the maxims of relevance and quality here.

[3] In terms of Sperber and Wilson's theory of relevance this would probably mean that humans always try to maximize the relevance of the information that is being processed; in other words, they try to improve their knowledge of the world as much as possible given the available resources (Sperber and Wilson, 1986). However, the idea of relevance will not be pursued any further in this chapter.

description, given the principle of minimal cooperative total effort.

## 3.4 Focus of attention

An important determinant of the ease with which an object is identified is its relative *salience* in the context of the domain at some point during the interaction. The concept of salience has a two-way relationship with the focus of attention of the participants. On the one hand, an object that is salient at some point can be said to attract the focus of attention of the participants. On the other hand, an object that is in some way in the focus of attention of the participants can be said to be more salient. There are various ways for something to gain salience; some have to do with the course of conversation, others do not (Lewis, 1979). In Lewis's view, some contextually determined salience ranking occurs that may change during the course of conversation.

In our opinion, there are at least three ways in which an object can become salient and/or part of the current focus of attention. First, an object can acquire an inherent salience if at some point during the interaction it stands out in the context. Secondly, an object may be salient either if it has been mentioned recently or is in some way related to an entity that has been mentioned earlier, or if the attention has been pulled toward it in some other way. Thirdly, an object may become salient if it is functionally relevant in the current context. If an object is salient at some point during the interaction, and the speaker wants to refer to this object, then he or she will generally need less information to do this, because there are less other competing (i.e., salient) objects from which the target object has to be distinguished.

### 3.4.1 Inherent salience

Objects that are salient within the domain of conversation attract attention[4]. What salience means for the identification of objects was shown by Treisman and Gelade. They found that if a target item differed from the irrelevant items with respect to a simple feature such as orientation or colour, observers could detect the target just as fast when it was presented in an array of 39 items as when it was presented in an array of 3 (Treisman and Gelade, 1980). This observation is known as the 'pop out' effect. In addition, research using eye movement tracking has shown that objects with a high information content, i.e., more recognizable objects, tend to be fixated upon longer (Mackworth and Morandi, 1967). This observation holds also for objects that are unfamiliar in a certain situation (Loftus and Mackworth, 1978). Hence, it seems reasonable to conclude that objects that differ with respect to their environment tend to capture more attention and, as a result, can be identified more easily.

Salience of an object can also arise from changes in the features of the object. Alerting mechanisms direct attention to any gross change in the environment after it has been detected (Glass and Holyoak, 1986). This means that if a visually detectable feature of an object changes, such as contrast or location, the attention is directed towards this

[4] Note that at some point during the interaction, the salience of objects may change because of changes in the domain of conversation.

object.

How salience of an object in a certain environment may influence the production of the expression to refer to this object and the effort to identify it was shown by Clark, Schreuder and Buttrick. In an experiment they carried out, listeners were able to identify objects on the basis of ambiguous references by choosing the object that was perceptually most salient (Clark, Schreuder and Buttrick, 1983).

To conclude, a salient object is easier to refer to, since it suffices to only use reduced information. A salient object is also easier for the listener to identify, since it differs from the environment. The following hypothesis, presented in the form of an instruction to the speaker, can be derived from the literature discussed above:

**Hypothesis 1**

*'If the target object is inherently salient within the domain of conversation, use reduced information.'*

### 3.4.2 Current focus of attention

When talking about focus of attention, a clear distinction has to be made between the focus of attention within the dialogue and the focus of attention within the domain of conversation. Research about the focus of attention within the dialogue has centred around the possibilities for using pronominal expressions to refer to an object that has been mentioned recently. It is well known from the literature that the current explicit or implicit focus of attention in the discourse may influence reference to objects (Grosz, 1977; Grosz and Sidner, 1986). One of the main findings of Grosz and Sidner is that pronominal reference in dialogues is only used to refer to entities that are part of the current *explicit focus of attention*, i.e., entities that have been mentioned very recently in the discourse (e.g., 'the book..., it...'). The *implicit focus of attention* plays a role if an entity that has just been mentioned has strong associations with other entities. In those cases, the associated entity can be referred to by means of a definite expression (e.g., 'the book..., the author...').

It can be argued that the current focus of attention within the dialogue consists of a collection of features of the entity that has been referred to recently (the explicit focus), possibly supplemented by some features of related entities (the implicit focus). If we look at focus like this, we can observe that the speaker is allowed to omit the features in the current referring expression that have already been mentioned in the previous expression. A clear example of this is the use of type information. If all of the objects being referred to have the same type (e.g., a block) it is not necessary to convey this information in every single referential expression that is used. Grammatically, these reductions are treated as cases of ellipsis. Links with objects mentioned previously can also be expressed explicitly, e.g., in expressions such as 'the same one'. The case of pronominal reference to objects that are referred to repeatedly can be seen as the extreme case, where all features of the two entities are identical and only a pronominal 'place-filler' is necessary.

The focus of attention within the dialogue coexists with a focus of attention within the domain of conversation. Beside the inherent salience of objects that may attract attention, which was discussed in the previous section, there is also a more dynamic component of the focus of attention. This is the focus of attention that is continually established and changed during the course of the dialogue and the actions in the domain of conversation. This focus can be seen as a kind of spotlight that is controlled by the participants as the interaction unfolds. The counterpart in the domain of the explicit focus of attention in the dialogue is the object that has just been manipulated. In many cases, this object is also the last one mentioned in the dialogue. If such an object is referred to for the second time pronominal reference is possible.

We will call the counterpart in the domain of the implicit focus of attention in the dialogue the *spatial focus of attention* (see also chapter 2; Cremers, 1994). It can be argued that the objects that are located close to the one that has just been mentioned and/or manipulated are in the spatial focus of attention. Together with the object in explicit focus they form a *focus area*. If a speaker refers to an object that is located within the focus area, only the objects in the focus area have to be considered as alternative target objects. This usually means that the amount of information in the referential expression is reduced, which leads us to the following hypothesis:

**Hypothesis 2**

*'If the target object is located in the current focus area, use only information that distinguishes the object from other objects in the focus area.'*

### 3.4.3 Functional relevance

In the referential behaviour that was discussed in the previous section the focus area that had been established earlier in the dialogue was taken for granted. However, the speaker can also actively direct the focus of attention to a particular object even when the object is not located in the current focus area and use reduced reference. He can do this because expressing the goal or task in a task-oriented dialogue will often provide information about the identity of the objects that are involved. For instance, Clark et al. report on experimental findings that suggest that ambiguous references can be resolved by choosing the object that the addressee considers instrumental to the speaker's goal (Clark, Schreuder and Buttrick, 1983).

Other evidence for the influence of the task on the resolution of referring expressions was provided by Wright (1990). As a result of an experiment in which subjects were asked to describe a route on a map by referring to landmarks that were drawn on it, Wright identified nonlinguistic constraints that were present in the task which facilitated successful reference to these landmarks, among which some windmills. Subjects were able to identify the intended windmill on the basis of the expression 'the windmill' only, because they assumed it was the next windmill on the route the speaker was describing to them. Wright concluded that speakers actively deploy knowledge of focus constraints in their choice of referring expressions.

Implications of these findings for our task domain are that speakers can use reduced reference to refer to objects irrespective of their location within the domain by taking the pre- or post-conditions of the action into account. The intended result of this strategy is that the focus of attention will only be directed at objects that are suitable for use in carrying out the action. These objects can be said to be in the *functional focus of attention* (see also Cremers, 1994). This leads to the following hypothesis:

**Hypothesis 3:**
*'Use only information that distinguishes the target object from other objects that would also be suitable for use in carrying out the current action.'*

## 3.5 Features in the description

In the previous section we have described what the effect of reducing the focus space is on the number of features that have to be used in referential expressions. A conclusion from this is that the smaller the space that has to be taken into consideration, relatively the less features have to be used. In this section we will try to describe which features, given the focus space, speakers prefer to use to refer to a target object.

In general, a speaker's referential expression indicating some object in the environment is a function of what alternative objects there are in the context of reference (Olson, 1970). Speakers try to choose the descriptive content that distinguishes the target object from the surrounding ones most effectively. If there are two distinguishing features that are equally powerful, usually the speaker chooses the one that is most salient (Herrmann, 1983).

From our perspective salience is not the predominant criterion for choosing a particular feature. Speakers have the choice to use either absolute or relative features to refer to a certain object. From the principle of minimal total cooperative effort the prediction can be made that speakers have a preference for using absolute features, since to produce and understand those features no other objects than the target object have to be taken into account. This implies for the speaker that only one object has to be described instead of two or more, and for the addressee that only one object has to be identified. Hence, we would expect that both speaker and addressee need to expend less effort when reference by means of absolute features is used.

However, sometimes uttering absolute features may cause problems from both a generation and an interpretation point of view, because the features are inherently difficult or because too many features are needed to distinguish the target object from other objects. Compare, for instance, the following utterances: 'het blok dat zich bevindt op de coördinaten 318, 248' ('the block that is located at the coordinates 318, 248') and 'het blok naast het grote blauwe blok' ('the block next to the large blue block'). In those cases it may be more efficient to (also) use relative features, since it may reduce the total amount of collaborative effort required to achieve the goal of the common knowledge that the target object has been identified. The point at which a speaker will shift from using absolute features to using relative features is a complicated matter which should be investigated empirically. These considerations lead us to the following hypothesis:

**Hypothesis 4:**
*'Use absolute features as much as possible and use relative features only if necessary.'*

If relative features are used, both speaker and addressee should be aware of the implicit or explicit relatum that should be chosen from the potential relata. From a language production point of view, it takes less effort to use an implicit relatum, since in that case the relatum does not have to be expressed. If there is no possibility for using an implicit relatum, an explicit relatum has to be chosen. This leads to a process of *recursion*: in order to refer to an object, some other object has to be referred to. If we apply the principle of minimal total cooperative effort again, we can predict that the chosen relatum will be an object that is relatively easy to identify. The hypothesis related to this observation is:

**Hypothesis 5:**
*'If an explicit relatum is needed for referring to the target object, choose as relatum an object that is in the focus of attention.'*

Probably the object that can be identified most easily is the object that was mentioned most recently, in other words, the object in the current explicit focus of attention. If the object in explicit focus is used as a relatum, it can be referred to by means of a pronominal expression. This results in a reduction of the number of words in the referential expression. If the target object is located close to an inherently salient object, this object can be chosen as a relatum. However, in that case pronominal reference is not possible.

## 3.6 Empirical setup

In order to find empirical evidence for the hypotheses that were formulated in sections 3.4 and 3.5, we carried out an empirical study during which two participants were asked to perform a specific task in a shared domain of conversation. The empirical situation is depicted in Figure 4 and can be described as follows.

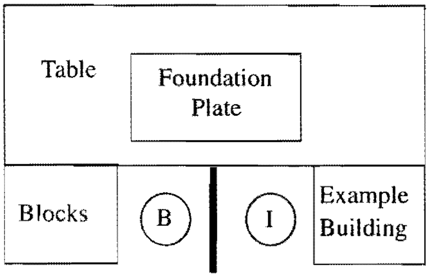


Figure 4 Empirical configuration (top view), B = builder, I = Instructor

Two participants were seated side by side at a table, but were separated by a screen. To avoid other communication than by spoken language and gesturing, only their hands were visible to one another, and only when placed on top of the table. One of the participants (the instructor, I) was told to instruct the other (the builder, B) in rebuilding a block building on a green toy foundation plate, located on top of the table such that the building would become a replica of the example building visible only to the instructor. Both participants were allowed to observe the building domain, to talk about it, and to gesticulate in it, but only the builder was allowed to manipulate blocks. The building consisted of blocks of one of four different colours (red, green, blue and yellow), three sizes (small, medium, large) and four shapes (square, bar, convex, concave).[5] Schematic pictures of the 29 blocks that were involved in the building sessions are provided in Figure 5. These objects were chosen because we wanted objects that were simple and non-figurative, in order to avoid extensive reasoning on domain specific knowledge by the participants.

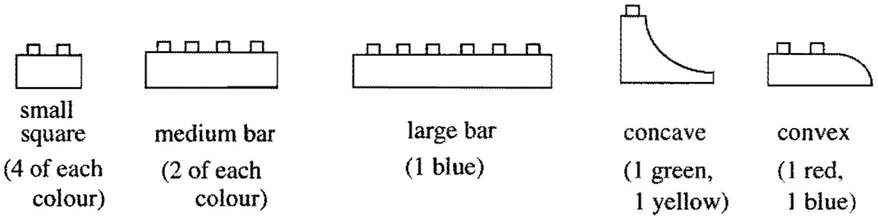


Figure 5 Types, numbers and colours of the blocks used in the study (side view).

Ten pairs of Dutch subjects participated in the empirical study. Half of the subjects was male and half female, and their ages varied from 20 to 60 years. The 10 building sessions were recorded on video-tape and the spoken communication was transcribed. The dialogues that occurred during the sessions were similar to Grosz's task dialogues (Grosz, 1977).

## 3.7 Criteria for determining the focus area

In our domain the spatial focus of attention is the predominant type of focus, since the nature of the task calls for the instructor to spatially scan the domain to look for parts of the block building that should be altered. Before we can talk about the features needed to refer to objects that are located in the current focus area, we first have to define the criteria for deciding whether an object is located within this area in the context of our empirical domain.

### 3.7.1 In focus

In our domain we can distinguish five indicators that determine if an object is located in the current focus area. Occurring indicators are either domain-related or linguistic crite-

[5] In fact, the blocks were samples of the DUPLO®-series of LEGO®.

ria, or combinations of both types.[6]

**Domain-related indicators for objects within the focus area**

- the target object is located adjacent (or relatively close) to the previous target object
- the target object is part of a set of objects that has been indicated in a previous utterance and identified by the partner (e.g., 'the group of blocks on the left')

**Linguistic indicators for objects within the focus area**

- a relatum *which is the previous target object* is used in the referential expression
- a definite expression is used, which indicates that the object is easy to identify
- linguistic markers that indicate to stay at the same location or that the (sub-)task has not been finished yet are used (e.g., 'here', 'we still have to...')

Example (4) illustrates the use of a referential expression to refer to an object within the focus area.[7] In this example a large and a small yellow block and a small blue block are all stacked on top of a red block that is mounted directly onto the foundation plate.

(4) I: Dit (raakt grote en kleine gele, kleine blauwe aan) moet er allemaal af.
B: (pakt grote en kleine gele, kleine blauwe vast)
I: ...(1.9) Blijft alleen <u>die rode op de grond</u> staan.
B: Ja ja. (haalt grote kleine gele, kleine blauwe eraf)

*I: These (touches large and small yellow one, small blue one) should all be removed.*
*B: (grips large and small yellow one, small blue one)*
*I: ...(1.9) Only <u>the red one</u> stays <u>on the ground</u>.*
*B: Yes yes. (removes large and small yellow ones, small blue one)*

In this example, 'die rode op de grond' ('the red one ... on the ground') was located in the vicinity of the large and small yellow ones and the small blue one. The referring expression is ambiguous within the domain, since at least one more red block was located at the foundation plate. Also, the definite expression 'de' ('the') is used. Furthermore, the uses of 'blijft' ('stays') and 'alleen' ('only') probably suggest that the total subtask has not been carried out yet, since they express a restriction to the number of blocks that have to be removed.

[6] In the list of criteria no task-oriented indicators are added. The possibility exists that the addressee is aware of the fact that the (sub-)task at hand is not finished yet, and that therefore the referential act is probably used to refer to an object within the current focus area. In our type of task this effect did not seem to be very prevalent, because the specific details with respect to the performance of the task were not prescribed. Task-oriented effects on the choice of references have been treated in depth by Grosz (1977).

[7] Comments by the transcriber about actions that were carried out are added between brackets in all examples.

### 3.7.2 Out of focus

If the target object is not located in the current focus area, a focus transition has to take place. Speakers may signal this transition explicitly by indicating the next focus area (e.g., 'let's go to the upper right part now'). If it is clear that the addressee has understood the nature of the transition, the next target object can be considered to be in focus. However, if no explicit indication is given, the referring expression itself should include enough information to identify the target object. Criteria that indicate that the target object is located outside of the current focus area are listed below. The domain-related indicators are complementary to those formulated earlier for objects within the focus area. The linguistic indicators are only partly complementary.

**Domain-related indicators for objects outside of the focus area**

- the target object is located relatively far from the previous target object (and certainly not adjacent to it)
- the target object is not part of the set of objects that were mentioned last

**Linguistic indicators for objects outside of the focus area**

- a relatum is used in the referential expression that is not the previous target object, but an inherently salient object
- an indefinite expression is used to indicate that the object is not easy to identify
- linguistic markers are used that indicate to move to another location or that the previous task or subtask has already been finished (e.g., 'let's move to the right', 'that part is ready')

In example (5) a focus transition to a new focus area is illustrated.

(5) B: Zo? (plaatst kleine blauwe)
I: Ja, ...(1.5) ja. ...(1.4) Nou, en -- Even kijken. Dan zie je op zeker moment, een beetje aan de noordkant, zie je <u>een groen blokje</u>.

*B: Like this? (places small blue one)*
*I: Yes, ...(1.5) yes. ...(1.4) Well, and -- Let's see. Then at a certain moment you see, a bit to the north side, you see <u>a green block.</u>*

In this example, the target object was located relatively far from the previous target object, and was not a part of some set of blocks introduced previously. Also, an indefinite referring expression is used: 'een' ('a'). Finally, a linguistic marker for a focus transition is given: 'een beetje aan de noordkant' ('a bit to the north side').

## 3.8 Results

During the execution of the task that was explained in section 3.6, subjects used a total of 665 referential acts. Of these references, 145 were first references to objects located in the domain of conversation. In the following we will report the results of analysing these first referential acts in terms of the hypotheses and criteria that were formulated in sections 3.4, 3.5 and 3.7.

### 3.8.1 In focus or out of focus?

Based on the criteria formulated in section 3.7 it was possible to make a distinction between objects that were in focus at the time of the utterance and those that were not. An object was determined to be part of the focus area if it satisfied one or more of the domain-related criteria. The linguistic criteria were only consulted if there was any doubt on the basis of the domain-related criteria. In total 99 (68%) of the 145 objects were located within the current focus area and 46 (32%) were located outside of the current focus area.

With respect to the domain-related criteria, we found that of the 99 objects that were determined to be located within the focus area, 80 were located relatively close to the previous object, and 9 were part of a set introduced previously. Of the 46 objects that were determined to be out of focus, 42 were located relatively far from the previous object and were never part of a set mentioned earlier. In 14 cases (10%) it was not possible to decide whether an object was in focus or not on the basis of the domain-related criteria alone. In 10 of these cases it was determined that the objects were in focus, and in 4 cases that they were out of focus.

Nine of the 10 objects that had been difficult to classifiy as being in focus had no direct contact with the previous object; however, they were objects closest to the previous object that had met the used description. In 6 of these 9 cases additional linguistic evidence for 'in focus' was provided. In 3 of these cases, relata were used that were previous target objects. In the other 3 cases as well as in the one remaining case linguistic markers were used that indicated that the task or subtask had not been finished yet. In all cases definite expressions were used.

Of the 4 doubtful objects that were decided to be out of focus, in 2 cases a linguistic marker was used that indicated that the previous task or subtask had been finished. In the other 2 cases a linguistic marker was used that indicated to move to another location.

### 3.8.2 Hypothesis 1

*1. 'If the target object is inherently salient within the domain of conversation, use reduced information.'*

In the domain, the concave and convex types were introduced to include inherent salience. Only 2 of each of both types appeared in the building. In total 10 first references were made to any of these objects, such as 'the green slide' to refer to a concave type and 'the half rounded one' to refer to a convex one. In 4 of these cases a colour feature plus pointing gesture were used, in 3 cases colour and shape were mentioned, in 2 cases demonstratives plus pointing were used, and in the remaining case only the shape was mentioned. Speakers did not use reduced reference as a result of salience in any of these cases. Either the referential acts were unambiguous within the whole domain or it was clear from the current spatial and/or functional focus of attention which object was being indicated. Therefore, it appeared not to be possible to find empirical evidence for the hypothesis. We will come back to this when discussing the results in section 3.9.

### 3.8.3 Hypotheses 2 and 3

2. *'If the target object is located in the current focus area, use only information that distinguishes the object from other objects in the focus area.'*
3. *'Use only information that distinguishes the target object from other objects that would also be suitable for use in carrying out the current action.'*

The existence of both a spatial and a functional focus of attention could clearly be observed during the interaction. The functional information that was made use of was related to the four basic operations that the participants were expected to carry out, namely, to remove an object from the domain, to move it within the domain, to leave it laying at the same location or to use it as a relatum.

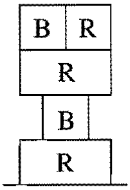
The total number of occurrences of reduction as a result of both the spatial and the functional focus of attention was 27 (27% of the 99 objects in focus). In all of these cases speakers used reduced information without pointing although other objects that met the description were present within the domain. This suggests that the expectations as they were formulated in hypotheses 2 and 3 were not met in the sense that participants used reduced information whenever this was possible. However, the fact that they used reduced information in 27% of the possible cases indicates that they were aware of the current focus of attention, and that they assumed that their partners were aware of it as well. We will come back to this point in section 3.9.

In 20 of the cases where reduction took place (74% of 27) it was not possible to divide the spatial and functional focus of attention, since the target objects were both close to the object mentioned previously and functionally relevant. In 17 of these cases (85% of 20) recognition of just the spatial focus of attention sufficed to identify the target object, since there was just one object present within the current focus area that met the description. In the 3 remaining cases (15% of 20) recognition of the functional focus of attention was essential, because there was more than one object present within the focus area that met the description but there was only one object that was functionally relevant as well.

In the 7 cases where a distinction could be made between spatial and functional focus (26% of 27) only recognition of the spatial focus of attention was possible. These were exactly the cases where the speaker was indicating that the object should stay at the same place.

An example of a reduction where both the spatial and the functional focus of attention are involved is (6). In this example a red block is located at the bottom with a blue block on top of it. A red block is placed on top of this blue block, and on top of this red one are a blue block and a red block. A schematic picture of this configuration is provided on the right of the dialogue fragment.

(6) I: Alles wat achter die gele steen staat (wijst) dat mag weg.
B: (verwijdert kleine groene, kleine blauwe)
I: Dus <u>die rode steen</u> gaat eraf, en alles wat erop staat.
B: Even kijken, dus dit moet er dus allemaal af
(haalt twee kleine blauwe, kleine rode en grote rode eraf)
I: Ja.
B: En dit ook he? (haalt grote rode eraf)
I: Ja.
B: Ja.



I: *Everything lying behind that yellow stone (points) can be removed.*
B: *(removes small green one, small blue one)*
I: *So <u>the red stone</u> has to be removed, and everything on top of it.*
B: *Let's see, so all of these should be removed (removes two small blue ones, small and large red one)*
I: *Yes.*
B: *And this one too huh? (removes large red one)*
I: *Yes.*
B: *Yes.*

'Die rode steen' ('the red stone'), which is a reduced expression, is used to refer to one of the red blocks. However, three red blocks are located close to the 'gele steen' ('yellow stone'), so the spatial focus of attention does not provide enough disambiguating information. The additional information that 'alles wat erop staat' ('everything on top of it') should be removed too provides a solution to this problem. Two of the three red blocks that are present have other blocks on top of them, but the one at the bottom has the largest number of blocks on top of it, so this is functionally the most relevant object. This is actually how the addressee understood the expression, which happened to be the right decision.

### 3.8.4 Hypothesis 4

4. *'Use absolute features as much as possible and use relative features only if necessary.'* Of the total amount of 145 first referential acts that occurred in the dialogues, 82 (57%) just included absolute features. In 5 (3%) of the cases only relative features were used. The difference between these numbers is highly significant ($\chi^2$=68.16, $p<0.001$). This seems to suggest that speakers have a clear preference for using absolute features over relative features. In 34 cases (23%) combinations of relative and absolute features were used. Beside relative and absolute features, demonstrative expressions accompanied by a pointing gesture[8] were also used. This was done in 24 (17%) of the cases. In Figure 6 the percentages of absolute features, relative features and demonstratives used are shown.

[8] There appeared to be personal preferences for using gestures to refer to objects. Three out of 10 instructors did not use any gestures at all. This fact may have influenced the percentage provided here.

### 3.8.5 Hypothesis 5

*5. 'If an explicit relatum is needed for referring to the target object, choose as relatum an object that is in the focus of attention.'*

As can be concluded from Figure 6, a relative feature occurred in a total number of 39 (26%) referential acts. In 19 of these 39 occurrences (49%) an explicit relatum was used. Either participants or other objects or both were used as explicit relata. In 4 (21%) cases one or two of the participants were mentioned as relatum, in 13 (68%) cases some object in the domain served as explicit relatum, and in 2 (11%) cases both a participant and an object were explicit relata.

Participants are always in the focus of attention, because their perspective always has to be taken into account by the speaker while formulating the referential expression. This means that in the 4 cases where a participant was used as a relatum, the relatum was in the focus of attention.

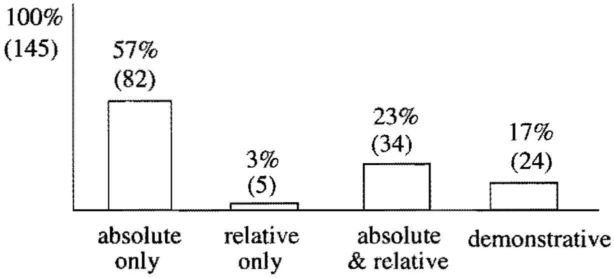


Figure 6 Percentages of features used as reference to objects (absolute, relative, absolute and relative, demonstratives).

Objects can be considered to be in the focus of attention if they have been mentioned previously (explicit focus of attention), are located in the current focus area (spatial focus of attention) or are inherently salient. The functional focus of attention does not apply here, because a relatum that is needed for referring to a target object is never involved in the action that should be carried out. If some domain object was used as a relatum (in 15 cases), in most cases (10, 67%) this object was in explicit focus. In one case (6%) the relatum was located in the current focus area. In the 4 remaining cases (27%) the relatum was either inherently salient or a unique object within the domain.

These results show that the relatum was in focus in *all* cases where an explicit relatum (either a participant or an object) was used. Examples (7)-(9) show, subsequently, the uses of a participant, an object in explicit focus, and a salient object serving as a relatum.

(7) I: O en de rode die <u>u</u> nou pakt,
die rode blijft ook zitten zie ik
B: Die blijft ook zitten.

I: *Oh and the red one that you're picking up right now,*  
*that red one stays there too I notice*  
*B: OK, I'll leave that one there.'*

(8) I: Dat gedeelte met die groene ronding  
B: Ja.  
I: die kan weg.  
B: Alleen de groene ronding?  
I: Die groene kan weg, en dan kan ik het pas zien.  
B: (verwijdert de groene ronde)  
I: Uh, dat gele aan de onderkant daarvan moet ook nog weg

*I: That rounded green part.*  
*B: Yes.*  
*I: that can be removed*  
*B: Only the rounded green one?*  
*I: I can't tell until that green one has been removed.*  
*B: (removes the rounded green one)*  
*I: Eh, that yellow one that was under it still has to be removed as well.*

(9) Die gele op die halve ronde, die kan d'r af.  
*The yellow one on top of that half rounded one can be removed.*

Example (7) illustrates how the speaker used his partner as a relatum by referring to the action she was executing at that moment ('die u nou pakt' ('that you're picking up right now')). In example (8), an explicit relatum is used in the form of a pronominal expression ('daar' ('it')), which is the last object mentioned before. Note that 'de groene ronding' ('the green rounding') could serve as a relatum, although at that point it had already been removed from the domain. This is an indication that objects that do not 'exist' any more can be used to refer to another object, and, *inter alia*, so can historical events (see chapter 1; Cremers, 1994). In example (9) an explicit relatum is used in the form of a definite description ('die halve ronde', 'the half rounded one') that was a salient object within the domain and had not been mentioned before.

## 3.9 Discussion

Some issues related to the results of the empirical study still remain unsolved or need to be discussed in further detail. These issues are particularly associated with the salience of objects, redundancy of information used in referential expressions, focus clashes between different types of focus, and the imbalance between the number of referential expressions used to refer to objects in and out of focus. These issues as well as some limitations to the present study will be discussed below.

### 3.9.1 Salience

In the results of the empirical study we were not able to find conclusive evidence for hypothesis 1, which stated that reduced information suffices to refer to salient objects. This negative result was probably due to the fact that the salient objects (i.e., the concave and convex objects) did not stand out enough in the domain. It would have been possible for the participants to only use the shape feature, which was the most salient feature, to refer to these objects. However, they did not do this systematically.

An explanation for the absence of reduced reference could be the *set effect.* This means that if a speaker has mentioned a feature over and over again, he will tend to keep on using it, even though it no longer has discriminating power (Herrmann, 1983). Indeed, in 7 of the 10 referential acts used to refer to concave or convex objects (70%), a colour feature was used although this was not needed for disambiguation. The major part of the preceding references included colour features as well. The participants were probably used to using colour and did not switch to shape if they encountered a salient object. Another possibility is that they experienced colour as being a salient feature that comes to mind first.

Although the results may suggest otherwise, we still think that salience is an important feature of objects that may even overrule the current spatial and/or functional focus of attention. To check this conviction tailored experiments should be designed.

### 3.9.2 Redundancy of information

Although we did not find conclusive evidence for hypotheses 2 and 3 in a strict sense, we found convincing empirical evidence for the fact that participants were making use of the focus of attention. They reduced the information in 27% of the 99 referential acts to refer to objects in focus. This means that in the remaining 73% they were using more information than was actually necessary according to the principle of minimal cooperative total effort.

One reason for this redundancy of information is the set effect that was already mentioned earlier (see Herrmann, 1983). Another possible reason for the redundancy is the *endophoric* redundancy, which is associated with the preceding part of the discourse (Pechmann, 1984). This means that speakers refer to an object by contrasting it to the one last focused on by the listener, thereby providing more information than is actually necessary. An example of such a sequence is: ‘the blue square, the red square’, used in a situation where no other squares are present in the immediate environment, so ‘the red one’ should provide enough information. Pechmann also assumes that a production problem could be involved. In those cases the speaker is simply not able to decide quickly enough which and how many of the possible features to use, so just starts naming features he or she considers to be appropriate. Pechmann finally argues that speakers deliberately give more information to help their hearers to find the target object. In our words, they place a relatively larger part of the cooperative effort at their own side of the scale. This can be explained by realizing that speakers probably give more information to avoid having to engage in an explanatory sub-dialogue in case the hearer has not understood the initial expression.

In our domain it is hard to decide which of the possible reasons for redundancy apply in a particular situation, since our empirical study was not as controlled as the ones Pechmann and Herrmann designed. However, there is strong evidence that the set effect plays an important role here, because colour is used in 95% (115 out of 121) of all information-containing utterances (i.e., all references except demonstratives).

### 3.9.3 Focus clashes

We already stated that it often is not possible to decide which of hypotheses 2 and 3 is being applied when a reduced expression is used. As a result of a clash of these two types of focus, miscommunications or uncertainties may occur. An example of an uncertainty is example (10), which was uttered in a situation where a small yellow block was placed on top of a large yellow block.

(10) I: Dan daar achter die gele d'r ook af pakken (raakt grote gele aan)
B: De grote gele?
I: Ja, de grote.

*I: Then behind that one also remove that yellow one (touches large yellow one)*
*B: The large yellow one?*
*I: Yes, the large one.*

In this example, a conflict occurs between the functional and the spatial focus of attention. Although I refers to a large yellow block and even touches it, B is not sure that I really intends to refer to that particular block, and consequently asks for confirmation. Most likely the reason for this uncertainty was the fact that both yellow blocks were in the spatial focus of attention, but only the small block was functionally relevant as well, since it was the one that could be removed most easily.

In principle clashes between inherent salience and the spatial or functional focus of attention are possible too, but these did not occur in our data.

### 3.9.4 Imbalance of references to objects in focus and out of focus

The distribution of first references referring to objects within the focus area as opposed to objects out of the focus area turned out not to be balanced (68% in focus, 32% out of focus). In terms of the principle of minimal cooperative total effort there are two possible reasons for this imbalance.

In the first place, people may have a preference for referring to objects in focus, because the referential expression that is needed will generally be shorter, and the chance that only absolute features are needed will be larger.

The second reason is that there may be a preference for staying in the same focus area or even choosing the object that is directly connected to (i.e., touching) to the one mentioned previously. This preference is the result of a higher level general strategy to solve problems. When people are trying to solve a complicated problem, they tend to decompose this problem and first solve the parts before solving the whole (Thomas, 1974). In terms of the blockbuilding task this would mean that participants first finish a part of the building (which is probably also the current focus area), and then choose a new part until the whole building has been completed. This strategy takes less effort than the alternative strategy which suggests to move to another focus area after every referential act. The problem of having to return to a previous focus area because a part of it has not been revised yet is also avoided.

Following the general problem solving strategy, participants prefer to choose an object within the current focus area. Exactly which object is chosen as the next target object is probably related to a lower level principle, the *principle of connectivity,* that was formulated by Levelt (1982). Subjects applied this principle when asked to describe spatial-grid-like networks. They chose as the next node to be described, wherever possible, one that had a direct connection to the current node. Levelt states that the principle of connectivity is a general ordering principle in perception and memory. However, he does not explain why this is the case. This lower level process probably works in the same way as the higher level problem solving strategy. Speakers probably choose the object closest to the previous one, in order to use less effort than would be needed to 'switch' to some object located further away (but still within the focus area). They also try to keep track of what they have been doing in order not to forget an object, since in that case they would have to return to it later, probably even after already having left the current focus area.

By applying the problem solving strategy of using subgoals and the principle of connectivity, coherence in the discourse may arise. If a focus transition marker is used, it may be relative with respect to the previous focus area (e.g., 'move further to the right'), and in this way connect the new discourse segment (and also the new focus area) to the previous one. Within a focus area, explicit connections can be expressed by using the previous target object as a relatum for the current one (e.g., 'the yellow block to the right of it'). However, participants may experience a sense of coherence even if coherence in the discourse is not created explicitly by expressing the relation between the previous and the current target object, because of the visual feedback they receive from the domain of conversation. For example, if no explicit relatum is used, participants can still *see* that the current target object is located close to the previous one, and may feel that the choice of the current target object is a coherent move in the interaction.

By using the term 'focus' for all types of focus that have been discussed in this chapter, we can state that, in our domain, focus is the main cause of coherence. We should however be careful not to extrapolate these findings to other domains of conversation too easily. On the one hand, in order to communicate about the present domain not much world knowledge was needed, so top-down coherence-establishing devices such as scripts and frames (see Brown and Yule, 1983) were not used. On the other hand, it may turn out that scripts and frames can be interpreted as devices that highlight certain entities in a particular context, hereby bringing these entities into 'focus'.

### 3.9.5 Limitations

The present study is limited in a number of ways. In the first place, we have focused on the descriptive content of the referential act, because this is the main part where information is localized that helps the addressee to identify the referent object. However, beside the descriptive content, determiners and gestures may also form part of the referential act.

Important information is expressed in the determiner that helps to carry out the identification; the information about the accessibility of the referent (Ariel, 1990) is especially useful here. It has been found that both in the discourse and in the physical domain Dutch speakers use 'dit/deze' ('this') to refer to objects with a relatively low accessibility and 'dat/die' ('that') to refer to objects with a relatively high accessibility (Kirsner, 1979). In our domain objects out of focus have a relatively low accessibility, while the accessibility of objects in focus is relatively high. The demonstratives 'dit'/ 'deze' and 'dat'/'die' were used accordingly (Piwek, Beun and Cremers (1995, 1996)).

Of course, important information can also be expressed by means of gestures. Not only can gestures help to identify a location, but they can also indicate, for example, shapes and sizes of objects (Knapp and Hall, 1992). In the referential acts we studied only pointing gestures were used in order to support the verbal information.

Also, we did not take into account the process of cooperatively building up to the agreement that a certain object is indeed the referent object. We assumed that just one referential act would suffice to achieve this. In reality this was not true, and sometimes more turns were needed[9], mainly at places where misunderstandings occurred. Main causes for miscommunication can be erroneous specificity, improper focus, wrong context or a bad analogy with another object (Goodman, 1986). In our data, 6 occurrences of confusions and/or miscommunications occurred (in 4% of the first references to objects in the domain). In one case the misunderstanding took place because the instructor provided wrong information. In all other (5) cases misunderstandings were in some way related to the focus of attention. In two cases the instructor probably assumed that the focus was still directed at a certain focus area and accordingly used reduced reference, which the builder failed to understand immediately. In two cases misunderstandings occurred at focus transitions, probably because it was not clear to the builder what the new focus area was going to be. One misunderstanding was the result of a focus clash that has already been discussed in section 3.9.3 and illustrated by example (10).

A final important limitation of this study is that we have only analysed referential behaviour in a blocks domain during a building task. In other types of domains and/or tasks the focus mechanisms and the choice of the types of features could turn out to be different from what we found. For example, in another type of task the functional focus may be more prevalent than was the case here. However, we claim that by choosing simple nonfigurative objects and a simple task, we were able to find basic characteristics underlying object reference.

[9] See chapter 4 (Cremers, 1995a) for a study of the cooperative process of reaching agreement that a certain object is the target object, using the same empirical data.

## 3.10 Conclusions

In this chapter, we have tried to describe what we think are basic principles underlying the choice of a particular type of referential act to refer to an object in a shared domain of conversation in which a task is carried out cooperatively. We have formulated the principle of minimal cooperative total effort, an extension of Clark and Wilkes-Gibbs's principle of minimal cooperative effort. We were able to formulate two consequences of this principle. First, speakers limit the number of potential alternative target objects by making use of the assumed focus of attention of their addressees. Second, speakers try to include as few objects as possible in the referential expression itself, either explicitly or implicitly. These two devices help, on the one hand, to keep the referential expression as short as possible, and, on the other hand, to limit the number of objects that have to be considered in order to find the target object. Thus, the principle of minimal cooperative total effort cuts both ways here. It takes less effort both for the speaker to utter the expression and for the addressee to identify the target object.

We were able to show that focus is not only a discourse-related phenomenon, but is also present in the domain of conversation. In both cases, if an object is in the current focus of attention, reduced information to refer to this object can be used. In our empirical study we found that speakers used reduced information in almost one-third of the cases where the target object was located in the focus area to refer to an object for the first time. Speakers also tried to avoid using explicit relative features. They only used these features if this was really necessary in order to avoid ambiguities. The relata that were used were always either objects in the current focus of attention or salient objects.

By relating the current target object to the previous object, either implicitly when the target object is located close to the previous object, or explicitly when the previous object is used as a relatum to indicate the target object, coherence in the interaction is established. Hence, the principle of minimal cooperative total effort contributes to a sense of coherence in the interaction.

The limitations of the present study are mainly due to the type of referential acts that were studied (first references with the emphasis on the descriptive content), and to the choice of domain and the task that was carried out. Future research should be broadened to include non-initial referential acts, other tasks and domains. Other modalities of communication, for instance typed communication, should also be investigated. Further, the cooperative process of establishing the common knowledge that the target object has been identified should be studied, since in general more than one turn is needed to reach this knowledge. A final point of interest is to experimentally 'prove' under more controlled circumstances that humans really make use of the types of focus we found.

# 4 The process of cooperative object reference in a shared domain

## *Abstract*

*If a participant in a dialogue refers to an object it usually takes several turns until both participants have mutually accepted that the right object has been identified. This process of cooperative object reference is investigated in dialogues that resulted from an empirical study that involved two participants carrying out a simple task in a shared domain.*

*Point of departure of the study is that the participants try to minimize effort in reaching mutual acceptance by taking into account the current spatial focus area, and by a preference for including absolute features (as opposed to relative features) in the referring expressions.*

*The results show that minimization of effort was indeed striven for by the participants. To refer to objects within the focus area fewer turns were needed and a simple type of noun phrase (i.e., the elementary noun phrase) was mainly used. At focus transitions more complex noun phrases and more refashionings (changes in noun phrases that had been used initially) were used. Furthermore, there was a preference for using absolute features in elementary noun phrases. Relative features were mainly used at focus transitions in the first parts of complex noun phrases.*

*It is concluded that the characteristics of a certain communicative situation (the available modalities of communication, the type of domain of communication, the types of objects) provide the participants with specific means to minimize effort.*

✳ This chapter is a slightly revised version of the paper that has been published as: Cremers, A.H.M. (1995a) The process of cooperative object reference in a shared domain. In: Fava, E. (ed.) *Speech acts and linguistic research: proceedings of the workshop, July 15-17, 1994. A participant symposium held during the First International Summer Institute in Cognitive Science. Center for Cognitive Science, State University of New York at Buffalo, Buffalo, USA.* Padova, edizioni nemo, pp. 139-153.

## 4.1 Introduction

The discussion that occurs between two people who cooperate in carrying out a task is called a task-oriented dialogue. If the participants in the dialogue have agreed that one of them is to be the instructor, and that the other is to execute the instructions, then this division of labour will influence the form and content of the dialogue. One important consequence is that the instructor must provide information that makes it clear to his partner which objects are involved in the task.

If the task is being carried out in a so-called shared domain, i.e., a domain that is both visually and physically accessible to both participants, then the participants can refer to an object that is chosen from the domain by means of referential acts. These acts consist of referring expressions (e.g., 'that red block on the right') and/or (pointing) gestures (indicated by '↗'). A referring expression may consist of a determiner (mainly demonstrative expressions) and the descriptive content. For instance, in 'that red block on the right', the determiner is 'that' and the descriptive content is 'red block on the right'. In this chapter, the focus will be placed on the descriptive content, since this is the part where the speaker actually provides features of the object to be identified (i.e., the target object) and this information is essential for successful object reference.

Cooperation between the instructor and the executor increases the likelihood that an efficient solution will be found and thereby that the task will be performed satisfactorily. This cooperativeness also plays a role in the choice of the descriptive content to refer to a certain target object. Generally, participants will try to minimize the total amount of cooperative effort that is necessary to come to the mutual acceptance that the target object has been identified. The participants can do this by minimizing the total number of words used to achieve mutual agreement (Clark and Wilkes-Gibbs, 1986). However, the expended effort is also reflected in the choice of the features to be included in the referring expression, as well as in the time it takes to actually identify the target object and to reach mutual acceptance (chapter 3; Cremers and Beun, 1995).

This chapter deals with the process of reaching mutual acceptance about the identification of objects located in a shared domain that are referred to for the first time during a task-oriented dialogue. Only first references are investigated, since these always imply direct reference to objects in the domain, so the effort to refer to and to identify the object can be determined by inspecting the specific characteristics of the domain at hand. It is assumed that during this process the participants try to minimize the total amount of cooperative effort expended. On the basis of the assumption of minimization, some predictions about the course of the process of mutual acceptance will be formulated and subsequently tested in a corpus of task-oriented dialogues that was collected during an empirical study.

Object reference is discussed in section 4.2 in terms of the minimization of cooperative effort as described by Clark and Wilkes-Gibbs (ibid.) and Cremers and Beun (ibid.). The cooperative process of object reference is described in section 4.3, again based on the work by Clark and Wilkes-Gibbs. Some predictions about the expected course of this process are then presented. In section 4.4 the empirical method that was designed to test these predictions is described. Empirical findings are presented in section 4.5 and discussed in section 4.6. Finally, some conclusions are drawn in section 4.7.

## 4.2 Object reference in a shared domain

The pragmatics of object reference in a shared domain can be said to be based on the *principle of minimal cooperative effort* which was formulated by Clark and Wilkes-Gibbs (1986) in the tradition of Grice's maxim of quantity (Grice, 1975). Clark and Wilkes-Gibbs stated that there is a trade-off between the length of the noun phrase that is uttered first to refer to an object and the total length of the possible additions or corrections to this utterance by the speaker or the partner. One assumed overall goal of participants in a dialogue is to minimize the amount of cooperative effort needed to reach the common agreement that the target object has been identified.

In chapter 3 Cremers and Beun (Cremers and Beun, 1995) extended Clark and Wilkes-Gibbs's principle by stressing that the non-linguistic effort to identify the target object should be included in the cooperative effort as well. So, their *principle of minimal total cooperative effort* states that there is a trade-off between the noun phrase that is uttered first and the possible additions, corrections *and actions* by the speaker or the partner. The term 'action' should be taken in the broadest sense here to include physical actions as well as perceptual and (non-linguistic) cognitive actions leading to the identification of the intended object.

It is not very clear how to measure the absolute amount of effort that is needed, on the one hand, to utter a referring expression and, on the other hand, to identify the intended target object. However, it is possible to look for strategies that participants apply to *reduce* effort. Cremers and Beun (1995) found two ways in which participants in their empirical study reduced the cooperative effort.

The first way was to limit the area in which the target object had to be sought. This can be achieved if each participant takes the other participant's current focus of attention within the shared domain into account. If a speaker assumes that his or her partner's focus of attention is directed at a certain sub-area of the domain, then generally he or she can use less information to refer to the target object if it is located within the focus area than if it is located outside of it. In addition, the partner has to consider less objects when trying to identify the target object if the target object is part of the current focus area than if it is located outside of the focus area.

The second way in which cooperative effort was reduced was to choose a referential act that limited the number of objects that first had to identified in order to be able to identify the target object. This means that speakers preferred to use what was called *absolute features* instead of *relative features* in the descriptive content of the referential act. Absolute features are features that can be produced and understood by just considering the target object (e.g., 'the red block'). Relative features can only be produced and understood after first having considered objects other than the target object. Relative features can be either implicit (e.g., 'the large block'), in which case the other objects are not mentioned, or explicit (e.g., 'the block next to the red block'), in which case the other objects are mentioned (in this example 'the red block').

The results of an empirical study by Cremers and Beun (ibid.) show that participants in a task-oriented dialogue actually do try to minimize their total cooperative effort in the two ways described above. The participants clearly used less information to refer to objects within the focus area than they used for objects outside of the focus area. They also used significantly more absolute features than relative features, particularly when referring to target objects located within the focus area.

## 4.3 The process of object reference

### 4.3.1 Referring as a collaborative process

The process of object reference generally involves more than just the utterance of a referring expression by a speaker followed by the identification of the target object by the hearer. The participants may need to take many turns speaking and gesturing after the initial reference has been uttered before the target object is identified and mutual acceptance is reached.

Clark and Wilkes-Gibbs (1986) have analyzed this recursive process of mutual acceptance. They carried out an experiment in which subjects were asked to instruct their partners to put a set of complicated 'Tangram' figures in the same order as the set they had received themselves. The participants in this experiment did not have a shared domain, so they could only use referring expressions to refer to the 'Tangram' figures. They could not gesture and there was no visual feedback available about the actions of the partner.

On the basis of an analysis of the referential expressions that were used by the subjects, Clark and Wilkes-Gibbs argued that the process of mutual acceptance, of which a simplified version is presented here (see Figure 7), may consist of three stages. In the first stage, a speaker initiates a reference by presenting a noun phrase. Six different types of noun phrases were distinguished in the data:[1]

1. *Elementary noun phrase*: a single tone group forming one single noun phrase (e.g., 'the red block').
2. *Episodic noun phrase*: two or more easily distinguished episodes or tone groups forming one single noun phrase (e.g., 'the large red one, that is lying over there ↗').
3. *Installment noun phrase*: a single noun phrase formed by two or more episodes, or installments, which are separated by interruptions of explicit acceptance of each installment by the partner (e.g., A: 'The large red one.'; B: 'Yeah.'; A: 'That is lying over there ↗.')
4. *Provisional noun phrase*: immediate expansion of an initial noun phrase in a new noun phrase (e.g., 'the one below, the red one').
5. *Dummy noun phrase*: a single noun phrase used as a stand-in until a more complete noun phrase can be used (e.g., 'the what shall I call it, the green slide').

[1] The examples are not Clark and Wilkes-Gibbs's; they are either taken from our collection of empirical dialogues or constructed in terms of the empirical task.

6. *Proxy noun phrase*: an incomplete noun phrase finished by the partner, who has some confidence that she knows what the speaker was trying to say (e.g., A: 'The large blue block lying in front of the ...; B: 'large red one?').

Clark and Wilkes-Gibbs claim that there is an order of preference in this list. Of these six types of noun phrases, the elementary noun phrases are the most preferred, and the proxy noun phrases the least. However, they fail to give conclusive arguments for this ordering. Also, they admit that although their data are consistent with it, they are hardly definitive. Further research is clearly needed to confirm this claim.

In the second stage of the process of mutual acceptance, the initial noun phrase may be refashioned by either the original speaker or the addressee. The most conspicuous types of refashioning are the *request for expansion* and the *rejection.* In a request for expansion the addressee asks the original speaker to give more information (e.g., as a response to the initial noun phrase 'the red block' the addressee may say 'uh...' or 'which one do you mean?'). In a rejection, the original speaker disapproves of the noun phrase that was used (directly, e.g., by uttering 'no', or indirectly by providing an alternative noun phrase, e.g., 'the one that just fell down?'). Recursion may take place in the refashioning stage, since it may take several turns before both speaker and addressee are satisfied with the reference.

In the third and final stage, the reference is concluded by accepting it mutually. Acceptance can be asserted explicitly (e.g., 'okay'), or presupposed by continuing on to the next contribution. If no refashioning is needed, only the first and the third stages are used, and this is then called a *basic exchange.*

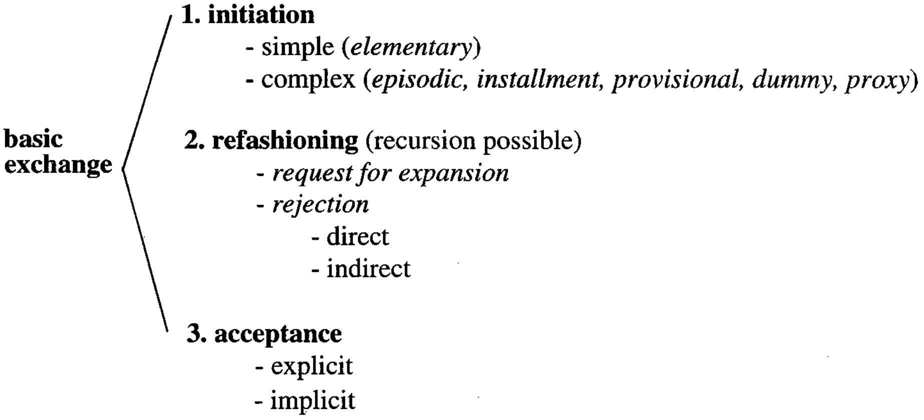

**basic exchange**
**1. initiation**
- simple (*elementary*)
- complex (*episodic, installment, provisional, dummy, proxy*)
**2. refashioning** (recursion possible)
- *request for expansion*
- *rejection*
- direct
- indirect
**3. acceptance**
- explicit
- implicit

Figure 7 Schematic overview of a simplified version of Clark and Wilkes-Gibbs's (1986) process of mutual acceptance.

### 4.3.2 Referring in a shared domain

Clark and Wilkes-Gibbs's (1986) data are based on references to relatively complex unique objects of which the location is irrelevant, since they are not part of a shared domain of conversation. They are inherently relatively hard to describe, and it is hard to find the relevant distinguishing features with other objects. In this chapter, the objects studied are inherently relatively simple and they are located in a shared domain in which they are not unique. The objects are inherently relatively easy to describe, but the speaker has to make sure that they are distinguished from other identical objects within the domain. To do this, absolute or relative location information can be used.

If we try to apply the principle of minimal cooperative total effort to the process of mutual acceptance in our domain, we can make some predictions about the course of this process. These predictions can be related to the two consequences of this principle that were discussed in section 4.2, i.e., firstly, speakers make use of the assumed current focus of attention by reducing the number of features included in the descriptive content, and secondly, they involve as few objects as possible in the referential act itself by preferring absolute features to relative features.

The first consequence can be paraphrased by saying that, generally, it takes less effort to refer to an object within the focus area than to an object outside of the focus area. In terms of the process of mutual acceptance, in our domain less effort can manifest itself in at least three ways.

The first indication of less effort is that fewer turns are needed to reach mutual acceptance. In the present context the prediction is that in the mutual acceptance process fewer turns are needed to refer to an object that is located within the current focus area than to refer to an object outside of the current focus area.

Secondly, less effort would also mean that speakers have a preference for uttering simple types of initial noun phrases since these are easier to produce and to understand. This leads to the prediction that to refer to an object in focus there is a preference for simple, i.e., elementary initial noun phrases, and to refer to an object out of focus there is a preference for complex (i.e., episodic, installment, provisional, dummy or proxy) initial noun phrases.

Finally, for the part of the acceptance process following the initial noun phrase, less effort would mean that participants try to avoid unnecessary refashioning. Therefore the prediction is that refashioning by means of requests for expansion or rejections will mainly occur at focus transitions.

The second consequence of minimal effort (concerning the preference for absolute features) can be paraphrased by saying that during the process of mutual acceptance, speakers will generally start by using absolute features, and will only use relative features if absolute features alone are not sufficient. This leads to the prediction that in elementary references, and in the initial parts of complex initial references, speakers have a preference for using absolute features. In the second part of complex references relatively more relative features are used.

## 4.4 Method of the empirical study

In order to test the predictions that were formulated in the previous section, an empirical study was carried out during which two participants were asked to perform a specific task in a shared domain of conversation. The situation in which the task was carried out is described in section 4.4.1 and is depicted in Figure 8. In section 4.4.2 the criteria that were used to decide whether a certain object was located within or outside of the focus area are listed.

### 4.4.1 Set-up of the study

Two participants were seated side by side at a table, but were separated by a screen. The situation was designed to limit communication to spoken language and gesturing. Therefore, only their hands were visible to one another, and then only when placed on top of the table. One of the participants (the instructor, I) was told to instruct the other (the builder, B) in rebuilding a block-building (the 'old' building) on a green toy foundation plate to become a replica of another block-building (the 'new' building) that was only visible to the instructor. Both participants were allowed to observe the building domain, to talk about it, and to gesticulate in it, but only the builder was allowed to manipulate blocks. Both buildings consisted of blocks of four different colours (red, green, blue and yellow), three sizes (small, medium, large) and four shapes (cube, bar, convex, concave).[2] This domain was chosen because we wanted the objects to be simple and non-figurative, in order to avoid extensive reasoning on domain specific knowledge by the participants.

Ten pairs of Dutch subjects participated in the study. Half of the subjects was male and half female, and their ages varied from 20 to 60 years. The 10 building sessions were recorded on video-tape and the spoken communication was transcribed. The dialogues that occurred during the building sessions were similar to Grosz's task dialogues (Grosz, 1977).

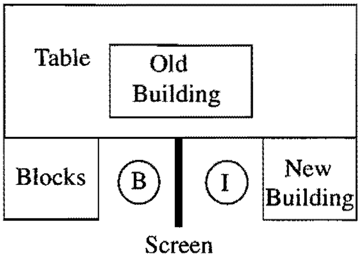


Figure 8 Situation of the empirical study, top view
(B = Builder, I = Instructor)

[2] In fact, DUPLO® -blocks by LEGO® were used

### 4.4.2 Criteria concerning the focus area

Before presenting the results of the study in terms of the predictions that were formulated in the previous section, it is necessary to define measurable criteria to determine whether these objects were located inside or outside of the current focus area. This could only be determined for a particular object if one or more of these criteria held. We will distinguish between domain-related and linguistic indicators. The domain-related indicators should be considered the most important criteria. In order to avoid circularity in the decision, the linguistic indicators were only checked if there was any doubt about the domain-related indicators.

**Target object in the current focus area**

*Domain-related indicators:*

- the target object is located adjacent (or relatively close) to the previous target object
- the target object is part of the area (e.g., 'the right upper part') or the larger whole of objects (e.g., 'the group of blocks on the left') that was indicated in a previous utterance and identified by the partner

*Linguistic indicators:*

- the referring expression is ambiguous in the context of the whole domain
- definite expressions are used to indicate that the object is easily identifiable
- linguistic markers are used that indicate either that the location should stay the same (e.g., 'here', 'there') or that the sub-task at hand is not finished yet (e.g., 'this part is almost ready').

**Target object not in the current focus area**

*Domain-related indicators:*

- the target object is located relatively far from the previous target object (and certainly not adjacent to it)
- the target object is not part of the most recently mentioned area or larger whole of objects

*Linguistic indicators:*

- the referring expression is non-ambiguous in the context of the whole domain
- ·indefinite expressions are used to indicate that the object is not easily identifiable · (non-)linguistic markers are used that indicate either that the (sub-)task at hand is finished or that another (sub-)task is going to start (e.g., 'this part is ready', 'let's see', pauses, hesitations) are used.

## 4.5 Results

In the dialogues that resulted from the conducted experiment, 145 processes of mutual acceptance occurred that led to the identification of some object that was located in the shared domain of conversation and was being referred to for the first time during the dialogue. According to the criteria described above, 99 (68%) of the 145 references were used to refer to an object that was located in the current focus area, and 46 (32%) were used to refer to an object outside of the current focus area.

### 4.5.1 Results concerning the focus area

#### 4.5.1.1 Number of turns

In order to test the prediction with respect to the number of turns both the linguistic and the non-linguistic acts were counted. Thus if a participant pointed at a certain object without saying anything, this was counted as a turn, even if this was done during the linguistic turn of the partner. The last turn that was counted on each occurrence was the one in which it was clear that the right object had been identified, and agreed upon by both parties. This agreement could be reached either explicitly or implicitly. In the case of explicit agreement, the last turn counted was the turn that settled this agreement, whether by means of language or by means of actions or gestures. In the case of implicit agreement, if no objections about the identification occurred later, then the last turn counted was the turn before the turn in which a new instruction was started.

A mean number of 2.4 turns (standard deviation = 0.8) in the dialogues was needed to refer to an object that was located within the focus area and to achieve mutual acceptance. To refer to an object outside of the focus area a mean number of 3.2 turns (standard deviation = 1.9) was needed. If we take $H_0$ to be that the mean difference between the number of turns in focus and out of focus is $\geq 0$, and $H_1$ that this difference is $<0$, then we can reject $H_0$ ($t=-3.478$, $p<0.0005$), so the difference between the means is significant.

An example of a sequence of three turns that occurred to establish mutual acceptance about the identification of an object within the focus area is:

(1) I: dat groene blokje dat daar rechts boven van die rode staat ('that green block standing on the right of and above that red one')
B: *(touches a small green block)*
I: ja ('yes').

An example of a sequence of 11 turns to identify an object outside of the current focus area is:

(2) I: Wat nou helemaal naar voren zit, daar zit die rode dwars. Daar moet ook nog een lange rode bovenop, maar die zit ook over de vier blauwe heen. ('What is laying in the front now, there lies that red one transversely. There should be a long red one on top of it, but that one also lies across the four blue ones.')
B: Uh
I: Dus helemaal vooraan zit een rode ('So totally in the front lies a red one')
B: Ja ('Yes')
I: Met acht gaatjes ('With eight holes')
B: Die ('that one' ) *touches a large red block*
I: O nee. ('Oh no')

B: Nee. ('No')
I: Vooraan ik bedoel, die.. ('In the front, I mean, that one ...')
B: O ja ach, ja ('Oh yes, ah yes') *touches another large red block* En daar moet een rode? ('And there should be a red one?')
I: Ja, dwars nu. ('Yes, transversely now')

#### 4.5.1.2 Initial noun phrases

The dialogues contained only three out of the six classes of initial noun phrases in Clark and Wilkes-Gibbs's classification, namely the elementary, episodic and provisional noun phrases. In Figure 9, the numbers of occurrences of initial elementary noun phrases and complex noun phrases (episodic and provisional) to refer to objects that were located either in or out of focus are depicted.

There is a significant dependency between on the one hand focus (either in focus or out of focus) and on the other hand, sentence type (either elementary noun phrases or complex noun phrases (both episodic and provisional)) ($\chi^2(1)=12.19$, $p<0.001$).

As is clear from the numbers of occurrences, the preference order of Clark and Wilkes-Gibbs does not hold here. Indeed, there is a clear preference for using elementary noun phrases, but the preference order of episodic and provisional noun phrases is interchanged. There were in total 22 occurrences of complex noun phrases, so if episodic noun phrases are preferred over provisional noun phrases, then the expected ratio of episodic:provisional is at least 12:10. It could be shown that this hypothesis does not hold ($\chi^2(1)=14.85$ $p<0.001$).

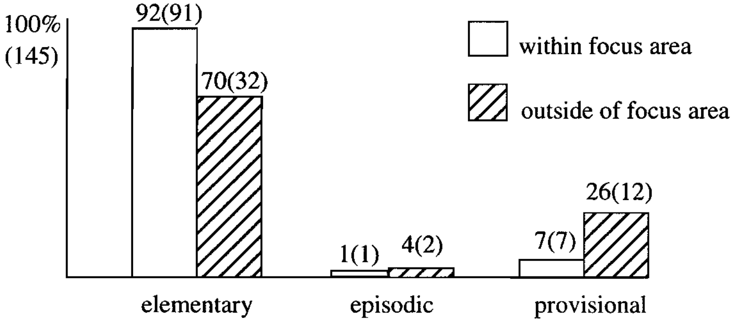


Figure 9 Occurrences of types of initial noun phrases (elementary, episodic and provisional) used to refer to objects within the focus area and outside of the focus area, in percentages(numbers). The percentages are calculated with respect to the total number of occurrences of each type of initial noun phrase within or outside of the focus area.

#### 4.5.1.3 Refashioning

In the dialogues, only six refashionings occurred, of which three were requests for expansion and three were rejections, for example:

(3) (request for expansion)
I: een groene vierkante (...) op het tweede niveau (...) ('a green square one (...) at the second level (...)')
B: Ja ('Yes') *touches a green block*
I: Nee ('No')

(4) (rejection)
I: Uh dan daar achter die gele d'r ook afpakken. ('Uh then there in the back also remove that yellow one')
B: De grote gele? ('The large yellow one?')

Of these six refashionings, five were used at focus transitions. These five refashionings were always due to obscurities with respect to the location of the target object. In the one case in which no focus transition occurred, the request for expansion was about a relative feature of the target object, namely the size:

(5) I: Die blauwe ook d'raf ('Remove also that blue one')
B: Die grote blauwe? ('That large blue one?').

Although the numbers are too small for a statistical test, they seem to show a tendency towards a confirmation of the prediction.

### 4.5.2 Results concerning the choice of features

In the dialogues, 123 elementary noun phrases were used that either contained absolute and/or relative features or just consisted of demonstrative determiners (e.g., 'this one'). In Figure 10, the distribution of absolute and relative features and demonstratives between elementary noun phrases, with and without accompanying gestures is shown.

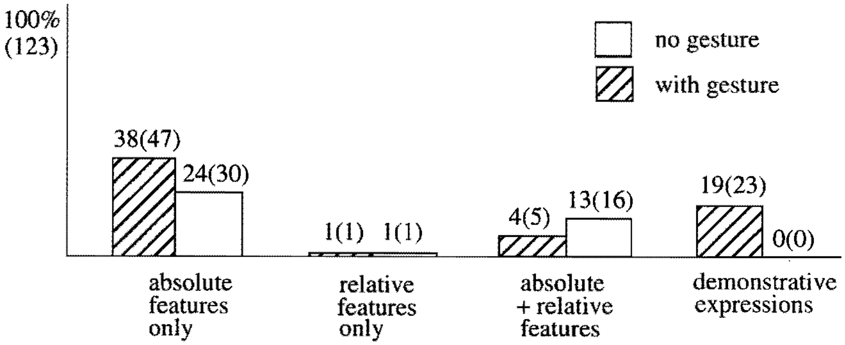


Figure 10 The choice of features in elementary noun phrases, with or without gestures, in percentages(numbers).

Significantly more absolute features only (62% (77)) than relative features only (2% (2)) were used in elementary noun phrases ($\chi^2(1)=71.20$, $p<0.001$). In 17% (21) of the cases, absolute and relative features were used together. Therefore in a total of 79% (98) of the cases absolute features occurred in elementary noun phrases. If we compare the uses of absolute features only and absolute and relative features, with and without gestures, we can see that if more features are used, then less gestures are needed ($\chi^2(2)=9.16$, $p<0.01$). The remaining cases are demonstrative expressions; these were always accompanied by gestures (19% (23)).

The numbers of absolute and/or relative features and demonstrative determiners that were used in the first (I) and the second (II) part of complex (provisional and episodic) noun phrases are shown in Figure 11. In the first part of a complex initial noun phrase there is no significant preference for using absolute features only (23% (5)) over relative features only (27% (6)). It appears that the difference between these types of features is not significant in the second part either, although there seems to be a trend to use more absolute features only (54% (12)) than relative features only (27% (6)).

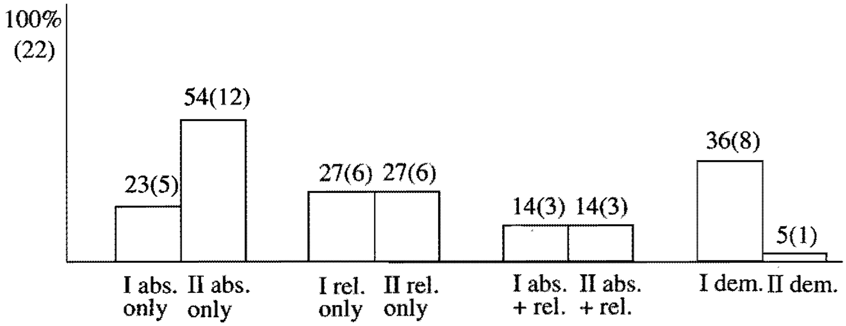


Figure 11 The choice of absolute and relative features, and demonstrative expressions in the first (I) and the second (II) part of provisional and episodic initial references, in percentages(numbers).

## 4.6 Discussion

### 4.6.1 Types of noun phrases

In the collected dialogues no installment, dummy or proxy noun phrases occurred. The absence of these noun phrases can be explained by comparing the type of references that were studied here with the type of references that were studied by Clark and Wilkes-Gibbs (1986). The difference can be understood by realizing that in their case it was far more difficult to find features to describe the objects than in our case. The speaker had to devise the appropriate features himself. In our case, the difficulty lay in choosing the most appropriate features in the current situation, and to do this only a limited number of features was needed.

Installments were probably not used because the descriptions were not complicated enough to have the participants feel the need for explicit acceptance in between. Dummy noun phrases were probably not used because the speaker did not have difficulties finding features to describe the current target object. Finally, proxy noun phrases were probably not used since the nature of the task did not allow the builder to predict which object would be the next target object. In addition, the objects have a lot of features in common, so the chance that the builder will add the right information in a situation where the description of the instructor is inadequate is small.

### 4.6.2 Features

The prediction about the choice of features that has been formulated in section 4.2 suggested that speakers have a preference for using absolute features over relative features. In the context of the empirical study this prediction would mean that in elementary references mainly absolute features were used. For the complex types of reference it would mean that in the first part mainly absolute features were used, and that in the second part relative references were used as well, but only if necessary.

The former of these predictions was confirmed by the results of the study. The latter was not confirmed, since speakers did not prefer to use absolute features in the first part of a complex reference. To find out why this was the case, the uses of relative features in the first part should be looked at more closely. In the dialogues, there were 6 occurrences of only relative features in the first part. In 4 out of these 6 cases (67%) a focus transition was taking place. The relative features were used here to indicate the new focus area in which the target object was located. An example of such a use is: 'Links onder, dat torentje met die afgeronde hoek.' ('On the left below, that little tower with the rounded corner'), in which 'links onder' ('on the left below') is an indication of the new focus area.

### 4.6.3 Identification failures

The cases where the hearer did not immediately identify the target object on the basis of the speaker's initial noun phrase can be considered identification failures. Apparently these identification failures occurred more often at focus transitions, since the mean number of turns to reach mutual acceptance was higher there. In his classification of reference identification failures, Goodman (1986) identified improper focus as one of the four causes of failure. In his terms, the failures that occurred in our material would be due to the fact that the speaker failed to distinguish the proper focus and did not notice the ambiguity for the addressee. These ambiguities occurred because the speaker failed to clearly indicate the location of the new target object or, in other words, the new focus area.

## 4.7 Summary and conclusions

In this chapter it was argued that the principle of minimal cooperative total effort applies during the process of mutual acceptance that a referential expression has been understood and that the target object has been identified. The processes of referring for the first time to objects that are located in a shared domain of conversation were analyzed based on two consequences of the principle of minimal cooperative total effort: 1. participants make use of the presence of a focus area, and 2. they have a preference for using absolute features.

During the process of reaching the mutual acceptance that the target object has been identified, the focus of attention appeared to have three effects. First, the mean number of turns needed to refer to an object located within the focus area was lower than the mean number of turns needed to refer to an object located out of the focus area. A second finding was that more elementary than complex (provisional and episodic) references were used to refer to objects in focus. The provisional and episodic references were used more often to refer to objects out of the focus area. Clark and Wilkes-Gibbs's (1986) preference order was not duplicated, since the order for provisional and episodic noun phrases was reversed. The third finding was a tendency to use refashionings at points where focus transitions occurred.

With respect to the choice of features, a clear preference for using absolute features was observed in elementary references. In provisional and episodic references there was no tendency to use absolute features in the first part. If relative features were used in the first part, this was done for indicating a new focus area. After having indicated the new focus area, absolute features were used in the second part to refer to the target object within this area.

A speaker can make optimal use of the principle of minimal cooperative total effort in several ways. In the first place the speaker can best choose as the next target object an object that is located within the current focus area, since in order to reach mutual acceptance about objects within the focus area, fewer turns are needed, elementary noun phrases usually suffice, and consequently absolute features can be used. Secondly, if it is unavoidable to shift to a new focus area, the speaker has to pay close attention to provide enough information, since focus transitions are potential sources of refashionings and identification failures.

The present study is limited to a situation in which both participants communicate verbally and gesturally about simple objects that are located in a task-oriented spatial domain to which both participants have access. Most probably the outcome of this study would have been very different if other modalities of communication or another type of domain had been used. For instance, if no gestures had been allowed, probably more turns or more complex noun phrases would have been needed to arrive at mutual acceptance. The same effect is expected if more complex objects had been used, such as the Tangram figures in Clark and Wilkes-Gibbs's (1986) study. Another spatial arrangement of the objects that had induced more necessary focus transitions would probably have resulted in more refashionings and identification failures. To validate these predictions more research is needed.

# 5 Object reference in task-oriented keyboard dialogues

## *Abstract*

*In the DenK project a multimodal interface is being developed which is suitable for graphical interaction as well as communication by means of natural language. For the design of this interface knowledge is needed about how humans refer to objects in a task-related environment, by means of natural language as well as gestures.*

*In this chapter some results of an experiment on referring behavior in task-related keyboard dialogues are reported, and compared to those of a preceding experiment on spoken dialogues. The differences that occurred between the two modalities were mainly related to the ease either to produce utterances, or to coordinate between using language, gesturing and inspecting the task domain or to change turns. These differences were all found to be based on the so-called principle of minimal cooperative total effort, i.e. within the limitations of the available modalities the participants tried to use as less effort as possible to, on the one hand, refer to a certain object, and, on the other hand, identify the object.*

*On the basis of the results some recommendations are provided for the design of a multimodal interface including the possibility of interaction by means of typed natural language.*

✳ This chapter has been published as:
Cremers, A.H.M. (1995b) Object reference during task-related terminal dialogues. In: Bunt, H., Beun, R-J. and Borghuis, T. (eds.) *Proceedings of the International Conference on Cooperative Multimodal Communication CMC/95, Eindhoven, May 24-26, 1995.* Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, pp. 115-128.

## 5.1 Introduction

In the so-called DenK project[1] (Ahn et al., 1995), a multimodal interface is being developed which is suitable for graphical interaction as well as communication by means of natural language. The DenK interface can be represented as a triangle as shown in Figure 12. The angles of this triangle stand for the *user*, the *domain* and the *cooperative assistant,* of which the latter two are components of the interface. The domain can be seen as the collection of objects represented on the screen and the relations between them. The cooperative assistant can be seen as the user's collocutor who is also able to perform actions in the domain. The user is allowed to point at objects in the domain or to manipulate them directly by means of some input device (e.g. a mouse). The user can also instruct the cooperative assistant by means of natural language to carry out certain actions in the domain, or ask questions about objects or events that play a role in the interaction.

If a user wants to ask questions or give instructions, it is important to make clear which objects are involved. In a multimodal interface the act of referring to objects can be performed by means of either natural language or pointing or a combination of the two. In any case, the user should take care to provide appropriate information for the system to be able to identify the intended object (the *target object*).

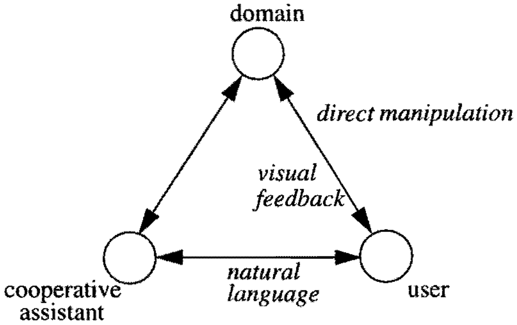


Figure 12 The DenK triangle

To equip the system with knowledge of how humans refer to objects in a ‘natural' situation, empirical research on this topic is needed. One of the most natural ways for humans to communicate is by means of speech. However, owing to technological limitations, most natural language systems only allow typed input. Unfortunately, it is not possible to extrapolate results from research on ‘natural' spoken dialogues to written dialogues. It has been shown that there are notable differences between the two modes of communication, in particular with respect to length and syntax (Hauptmann and Rudnicky, 1988),

[1] DenK stands for ‘Dialoogvoering en Kennisopbouw' in Dutch, which means ‘Dialogue Management and Knowledge Acquisition'. It is a joined research program of the universities of Tilburg and Eindhoven, and is partly financed by the Tilburg-Eindhoven Organisation for Inter-University Cooperation.

the speed and the planning of utterances, and the nature of the speech acts used (Oviatt and Cohen, 1991). For instance, more indirectness occurs in spoken dialogues than in keyboard dialogues (Beun and Bunt, 1987). In particular with respect to referential behavior it was found, when referring to objects for the first time, that in telephone (spoken) dialogues more requests for identification occur than in keyboard dialogues (Cohen, 1984). However, since this study dealt with telephone dialogues, only linguistic interaction was possible here. To enable conclusions about referential behavior in multimodal situations to be drawn, research on both spoken and typed dialogues is needed.

The referential behavior of participants in spoken task-related dialogues in a situation designed to mimic the DenK triangle has already been investigated in a previous study (chapter 3; Cremers and Beun, 1995). The present chapter deals with an empirical study on how humans refer to objects in a similar type of *keyboard* dialogue. The focus will lie on the type and amount of information humans use in referential expressions and on the use of gestures. The results of this study will be compared with findings from the previous research on spoken dialogues based on the differences between the two situations as represented in the DenK triangle.

In section 5.2 some results from the previous study on spoken dialogues are presented briefly. In section 5.3 some expectations are formulated about findings in a corpus of keyboard dialogues, based on the results obtained from the study on spoken dialogues and findings from the literature. In section 5.4 the results of checking the expectations in keyboard dialogues are presented and compared with the spoken dialogues. Finally, in section 5.5 the results will be discussed in the framework of DenK and some conclusions are formulated.

## 5.2 Referential behaviour in spoken dialogues

In a previous empirical study on spoken dialogues (chapter 3; Cremers and Beun, 1995) the referential behavior of ten pairs of participating subjects was investigated. The set-up of this empirical study is depicted in Figure 13a. The study was designed to mimic the triangular DenK paradigm and can be described as follows. Two participants were seated side by side at a table but separated by a screen. To prevent communication other than by speech and gesturing, only the hands of each were visible to the other participant and then only when placed on top of the table. One of the participants (the instructor) was told to instruct the other (the builder) in reconstructing a block building on a toy foundation plate, placed on top of the table, in accordance with an example provided. In this set-up the role of the instructor was similar to that of the user and the role of the builder was similar to that of the cooperative assistant in the DenK triangle. Both participants were allowed to observe the building domain, to talk about it and to gesticulate in it, but only the builder was allowed to manipulate blocks. A brief overview of the main results of this study will be given in the subsections which follow.

### 5.2.1 The principle of minimal cooperative total effort

In the experiment on spoken dialogues participants were found to adhere to the so-called *principle of minimal cooperative total effort.* This principle expresses the idea that together the participants try to say (Clark and Wilkes-Gibbs, 1986) and do (Cremers and Beun, 1995) as little as possible, but just enough to be able to reach mutual agreement that the target object has been identified. For the speaker this means that he or she will transfer the least possible information and also a particular type of information to refer to the target object, so that it allows the hearer to identify the object by having to consider as few objects as possible. Consequences of this principle in the spoken dialogues were related to the *choice of features* in the referential expressions and the *focus of attention* of the participants.

The first consequence was that, if possible, speakers preferred to use *absolute features* rather than *relative features.* Absolute features such as the physical feature 'red' are features that can be understood by considering only the target object. Relative features can only be understood by also considering other objects or persons that are present. Relative features may be either implicit or explicit. To understand implicit relative features, such as the physical feature 'large', other objects have to be considered. To understand explicit relative features, such as the location feature 'to the right of', other objects have to be identified in order to permit identification of the target object. Absolute features are consequently easier to understand than relative features.

The second consequence was that speakers used less information to refer to objects located in the area of the building domain that was in the current focus of attention of the participants than to those located outside of this area. As part of the task changes had to be made in several parts of the block building. If changes are being made in a particular part of the building the speaker can assume that the focus of attention of both himself or herself and his or her partner is directed at this area of the domain. For instance, participants used the referential expression 'the red block' to refer to the only red block within the current focus area, although many red blocks were present within the domain as a whole. Compared with the situation where the whole domain is taken into account, this means a reduction of words in the referential expression for the speaker, and fewer objects for the addressee to consider in order to find the target object.

Furthermore, it was found that participants in choosing the next object preferred to refer to an object that was in the current focus of attention. This resulted in a larger proportion of references to objects in focus (68%) than to objects out of focus (32%). In terms of minimal effort this could be explained as a strategy to make optimal use of the current focus area before moving on to the next one.

### 5.2.2 The process of object reference

In the spoken dialogues there were usually some turn-takings before the participants arrived at the common agreement that the target object had been identified. It was found in chapter 4 (Cremers, 1995a) that the number of turns needed was related to the focus of attention. To reach agreement on the identification of objects located within the current focus area fewer turns were needed than to refer to objects outside the current focus area (respectively 2.4 (s.d.=0.8) and 3.2 turns (s.d.=1.9)).

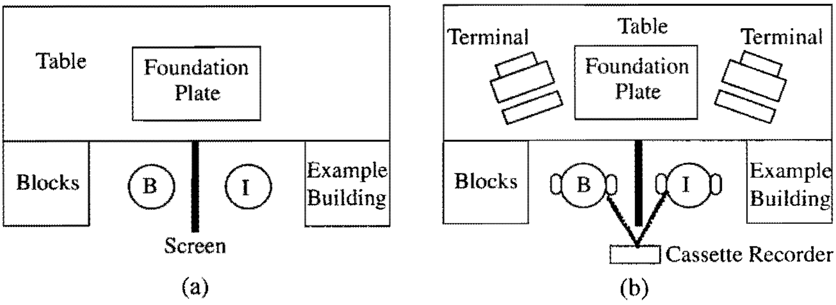


Figure 13 Empirical set-ups for (a) spoken dialogues and (b) keyboard dialogues.

## 5.3 Keyboard dialogues

In section 5.3.1 a description will first be given of the paradigm used for the study of multimodal keyboard dialogues, followed by an overview of the differences between this paradigm and the previous one on spoken dialogues that was discussed in section 5.2.

On the basis of the findings from the literature and from the preceding study on spoken dialogues some predictions for the outcome of this experiment will be formulated in section 5.3.2.

### 5.3.1 The empirical study

A second empirical study was carried out which was identical to that described in section 5.2, except for one important difference, namely that the participants communicated via keyboard and computer monitor instead of by speech. In the DenK triangle this means that the mode of communication between the user and the cooperative assistant is typed natural language. To prevent the participants from talking to each other instead of typing, they wore headphones to listen to some background music.[2] The empirical setup is depicted in Figure 13b.

The change from spoken communication to typed communication has some important expected consequences for the manner in which object reference can be carried out. First, the coordination between different modes of communication is expected to be different. In the spoken modality it is possible to speak and inspect the domain or point at objects in the domain at the same time. This is not possible in the keyboard situation. If a participant is typing, his or her attention is directed at the monitor and the keyboard, so that he or she can not see what is going on in the block-building process. Also, since his or her hands are busy typing, he cannot use them to point at objects in the domain.

A second consequence of the change from spoken to keyboard dialogues is that it is more difficult to take turns. To pass the turn on to the partner a participant had to

[2] This set-up served its purpose since the 10 pairs of subjects who participated never spoke to each other at any time during the experiment.

explicitly press a certain key. Only after he or she did so was the partner able to type. If a participant wanted to take his or her turn to type, he or she had to ask for it explicitly by means of a special key, and the partner had to acknowledge the switch of turn by pressing another key.[3]

Some expectations as to referential behavior in keyboard dialogues will now be formulated, based on the consequences of the use of typed communication instead of spoken communication.

## 5.3.2 Expectations for keyboard dialogues

### 5.3.2.1 Expectations about minimal effort

A general prediction with respect to keyboard dialogues that is an effect of the principle of minimal cooperative total effort is that it normally takes more effort to conduct a keyboard dialogue than a spoken dialogue, due to characteristics of the communicative modalities that are available. This difference in effort will be reflected in the length of referential expressions, the features chosen in the referential expressions and the use of gestures.

It is known from the literature (e.g., Oviatt and Cohen, 1991) that written dialogues generally take longer and contain fewer words than spoken dialogues. These results are also expected in the present study. The latter expectation also follows from the principle of minimal cooperative total effort. Since it takes more effort to type than to speak, fewer words will be used when typing. Written dialogues take more time than spoken dialogues but this increase would probably be even larger if more words were typed. However, the increase in time is not due only to the increase in effort. It can also be a consequence of the fact that participants do not feel as pressed for time as in spoken dialogues, so they take more time to formulate their utterances (Beun and Bunt, 1987).

With respect to the use of referential expressions in keyboard dialogues the participants are expected to try to utter the same information but use fewer words than in spoken dialogues. Probably also more gestures will be used, in order to compensate for the reduction in words.

With respect to the choice of features, the prediction is that, just as in spoken dialogues, participants will have a preference for using absolute features. There is no reason to assume that *more* absolute features will be used in keyboard dialogues, since the process of understanding a referential expression and identifying the referent is the same in both situations. An effect is, however, expected in the coordination of language and gestures. Since it is not possible to type and gesture at the same time, pointing gestures accompanied by demonstratives are expected to occur less often in keyboard dialogues.

As a result of a general reduction in words and an expected increase in the use of gestures, some of the features that were used in spoken dialogues will have to be replaced by gestures in typed dialogues. Tentative predictions are that absolute features

[3] If the participants had been allowed to type at the same time, this would have caused problems for them, especially since actions in the domain had to be monitored as well. In particular, the order of the turns and actions would have been less obvious.

containing information that can not be expressed very easily by gestures (e.g., color) will continue to be used, but that the rather verbose explicit relative features will be replaced by gestures.

The reduction of words as a result of the current focus of attention is expected to occur more often in keyboard dialogues than in spoken dialogues. A reduction of words means less typing and therefore less effort on the part of the participant. However, since the coordination of typing and inspecting the domain at the same time is difficult in keyboard dialogues, it is expected that participants will easily loose track of the current focus area. This will probably result in a relatively smaller number of references to objects in focus than in the spoken dialogues.

#### 5.3.2.2 Expectations about the process of object reference

In the spoken dialogues it was very easy to react immediately to something the partner said, resulting in a mean number of turns of 2.7 before mutual agreement was reached that the target object had been identified. The prediction for keyboard dialogues is that the effort to take turn to type will be so large that in most of the dialogues hardly any verbal turn-takings will take place. First, this could mean that more information will be given in the first turn, to avoid having to use more verbal turns. Note that this expectation contradicts the expected general reduction of words in referential expressions in keyboard dialogues. A second possible consequence is that the reduction in verbal turns will be compensated for by an increase in non-verbal turns since there is no inherent difficulty in taking turns in gesturing during keyboard dialogues.

There could be a reason for a possible *in*crease in verbal turns as well. This increase could be a result of the occurrence of more *miscommunications* during the keyboard dialogues, although it is suggested in the literature (see Cohen, 1984) that this effect does not exist. A miscommunication is defined as an event whereby a wrong selection takes place before the right target object is identified. The expectation of an increase in miscommunications is a consequence of the expected decrease in words in keyboard dialogues. To correct the miscommunication and identify the right target object additional turns will be needed. However, if the expectation about giving more information in the first turn to avoid having to engage in tedious turn-takings is correct, an increase in miscommunications is not likely to occur.

Finally, it is not clear whether in the keyboard dialogues, as in the spoken dialogues, the number of turns to refer to objects in focus will be lower than those to refer to objects out of focus. In keyboard dialogues, where participants have to divide their attention between keyboard, screen and domain, it is harder for them to continue focusing their attention on the current focus area. This could mean that they will not succeed in benefiting from the focus area as much as the participants in spoken dialogues did. In other words, it is probable that no difference in the number of turns will occur between in focus and out of focus.

## 5.4 Results

In the keyboard dialogues a total number of 156 referential acts occurred, which is almost the same as the number of referential acts found in the spoken dialogues, namely 145. This result is not surprising since both experiments involved exactly the same task and the same objects.

Findings with respect to the principle of minimal cooperative total effort and the process to reach mutual agreement on identification will now be discussed, and compared with the spoken dialogues.

### 5.4.1 Results concerning minimal effort

#### 5.4.1.1 Length

**Length of the dialogues**

In the literature it has been stated that, generally speaking, fewer words are used and more time is needed in keyboard dialogues than in spoken dialogues (Beun and Bunt, 1987; Oviatt and Cohen, 1991). This was also found in the present study (see Table 1). The participants took a mean time of 12 minutes to complete the keyboard dialogues, during which time they used 189 words. It took the participants a mean time of only 4 minutes and 47 seconds to complete the spoken dialogues, but in that time they used 729 words.

Table 1 Mean lengths of keyboard and spoken dialogues.

| | **keyboard** | **spoken** |
|---|---|---|
| **mean length** | 12 min.<br>(189 words) | 4 min. 47 sec.<br>(729 words) |
| **mean length of language** | 7 min. 56 sec.<br>(0.4 words/sec.) | 4 min. 17 sec.<br>(2.8 words/sec.) |
| **mean length of actions** | 4 min. 4 sec.<br>(34% of total time) | 30 sec.<br>(10% of total time) |

However, not all of the time was devoted to typing or speaking. A part of the time was used to carry out actions as well. The actions carried out were both pointing actions and manipulations within the domain. In the keyboard dialogues, 7 minutes and 56 seconds were taken for the actual typing, which means that the typing rate was 0.4 words per second. In the spoken dialogues, 4 minutes and 17 seconds were used for speaking, which yields a speaking rate of 2.8 words per second.

The figures show that in keyboard dialogues a relatively large part of the time was devoted to actions only, namely 4 minutes and 4 seconds, which is 34% of the time. In spoken dialogues 30 seconds were used for performing actions only, and that is 10% of the total time.

The results show that, indeed, it takes more time to conduct a keyboard dialogue than a spoken dialogue, under exactly the same conditions. In fact, it takes exactly seven times longer to type a word than to utter it. Also, the amount of time spent on carrying out actions is different for the two types of dialogue. In keyboard dialogues over three times longer is spent carrying out actions than in spoken dialogues. Since the task in the two experiments was exactly the same, this result cannot be explained by a difference in manipulating objects in the domain. The dissimilarity is therefore probably due to an increase in the use of referential actions, i.e. pointing or other gestures to indicate an object in the domain.

### Length of the referential expressions

A more specific hypothesis is concerned with the length of referential acts used in keyboard and spoken dialogues. The prediction was that, since fewer words are used in keyboard dialogues than in spoken dialogues, the length of referential acts in keyboard dialogues would also be shorter. This prediction did not completely prove true. Although the mean number of content words (i.e. all words except the determiner) used in keyboard dialogues was 1.8 (s.d. = 2.53), compared to 2.2 (s.d. = 2.69) in spoken dialogues, this difference does not mean that most references in keyboard dialogues were shorter than in spoken dialogues. First, the standard deviations are too large to show a clear difference in length between the two types of dialogue. Second, similar percentages of all lengths of referential expressions occurred in both dialogues, except for the referential expressions of lengths 0 and 1 (Table 2). More content-less referential acts, i.e. gestures or demonstratives or combinations of these, occurred in keyboard than in spoken dialogues (keyboard: 46%, spoken: 15%). In contrast, fewer referential acts containing only one content word occurred (keyboard: 12%, spoken: 46%).

Table 2 Numbers of content words in referential acts.

| length | keyboard | spoken |
|---:|---:|---:|
| 0 | 46% | 15% |
| 1 | 12% | 46% |
| 2 | 15% | 16% |
| 3 | 9% | 2% |
| 4 | 3% | 6% |
| 5 | 3% | 6% |
| >5 | 12% | 9% |

These figures seem to indicate that at times when typists use gestures only, or gestures accompanied by a demonstrative expression, speakers use one feature, possibly accompanied by a gesture, and vice versa. Since no large reduction of words in referential expressions could be demonstrated, the total reduction of words in keyboard dialogues must be due to a reduction of words in the remaining part of the utterances, i.e. the part where the action to be carried out is expressed.

However, if we do not count the number of words in the referential expressions but the referential expressions in which features are used a clear difference can be found. In keyboard dialogues fewer features (either absolute or relative or both) were used than in spoken dialogues, namely in 56% and 85%, respectively, of the referential expressions (see Table 3). This result is mainly due to the fact that in keyboard dialogues far more gestures without any language were used than in spoken dialogues, namely in 44% and 4%, respectively, of the references. Contrary to expectations, no difference could be found with respect to the total number of gestures used in keyboard and spoken dialogues. In both types of dialogue the percentage was exactly the same, viz. 53%.

Table 3 Features and gestures used in keyboard and spoken dialogues.

| | **keyboard (156)** | | **spoken (145)** | |
|---|---|---|---|---|
| | **+gesture** | **-gesture** | **+gesture** | **-gesture** |
| **absolute** | 9 (6%) | 38 (24%) | 45 (31%) | 43 (30%) |
| **relative** | - - | 2 (1%) | 2 (1%) | 2 (1%) |
| **abs. & rel.** | 4 (3%) | 35 (22%) | 7 (5%) | 24 (17%) |
| **demonstrative** | - - | - - | 17 (12%) | - - |
| **gesture only** | 68 (44%) | - - | 5 (4%) | - - |
| **total** | 81 (53%) | 75 (47%) | 76 (53%) | 69 (47%) |

### 5.4.1.2 Features and gestures

**Preference for absolute features**

One of the findings relating to the principle of minimal cooperative total effort in keyboard dialogues is, not surprisingly, that participants do have a preference for using absolute features rather than relative features, as is shown in Table 3. Absolute features only were used in 47 cases (30%). In spoken dialogues absolute features only were used in 88 (61%) of the referential acts. The use of relative features was more or less the same in both types of dialogue, viz. two (1%) in keyboard dialogues and four (2%) in spoken dialogues. Also, combinations of absolute and relative features occurred equally often in keyboard and spoken dialogues, viz. 39 (25%) and 31 (22%), respectively.

At first sight it may seem surprising that fewer absolute features were used in keyboard dialogues than in spoken dialogues. This seems to weaken the principle of minimal cooperative total effort. The solution to this problem lies in the use of gestures. If we assume that the use of gestures only or gestures combined with demonstratives is a means to use less effort, then the figures for the choice of features in keyboard and spoken dialogues become very similar. For keyboard dialogues this would mean that the referential acts which involve the least effort are those in which gestures only are used plus those in which only absolute features are used. These two percentages add up to 74%. In spoken dialogues, summation of the numbers of referential acts by means of gestures only, gestures plus demonstratives and absolute features only amounts to 77%.

To summarize, participants both in keyboard and in spoken dialogues try to reduce effort by choosing particular features. However, the choice of features is different in both types of dialogue. In keyboard dialogues relatively more gestures only are used and in spoken dialogues relatively more absolute features only.

### Coordination of typing and gesturing

The expectation with respect to the coordination of typing and gesturing was that in keyboard dialogues fewer demonstratives accompanied by gestures would occur. This indeed turned out to be true. In keyboard dialogues no cases at all occurred, whereas in spoken dialogues this combination occurred in 17% of the cases. This difference could even be extended to the use of absolute features accompanied by gestures. In keyboard dialogues they were used in 6% of cases, whereas in spoken dialogues they occurred in 31% of cases. Relative features and combinations of absolute and relative features accompanied by gestures occurred equally often in keyboard as in spoken dialogues.

### Type of features and gestures

The prediction concerning continuation of the use of features that cannot be expressed by means of gestures proved correct. In both types of dialogues almost the same percentage of absolute color features was used (keyboard: 100%, spoken: 97% of the absolute features used). However, there was a difference in the use of absolute shape features (e.g. 'square'). In keyboard dialogues 46% of the absolute features contained shape information, whereas in spoken dialogues this was the case in only 17%. A possible explanation for this difference is the fact that in spoken dialogues absolute features were about 4 times more often accompanied by gestures than in keyboard dialogues (keyboard: +gesture 9%, -gesture 46%; spoken: +gesture 36%, -gesture 47%). Since the use of pointing gestures makes the use of shape information superfluous, this type of information is probably used less in keyboard dialogues. The feature 'color' is probably so salient that participants tend to keep on using it, even though the use of a pointing gesture makes it superfluous.

The use of relative features in both types of dialogues was almost the same (keyboard: 1%, spoken: 2%). Although the number of explicit relative features in keyboard dialogues was lower than in spoken dialogues (keyboard: 23%, spoken: 39%) no clear difference was found. However, there was a difference in relative features that were used to refer to locations within the domain. If a location in the domain is indicated this generally takes relatively more words than if only physical features of objects are mentioned. It could be shown that in spoken dialogues more relative features were used to refer to locations (91% of the relative features used) than in keyboard dialogues (68%). This suggests that participants in keyboard dialogues tend to avoid these relatively long expressions, and probably point instead.

#### 5.4.1.3 Focus of attention

In the keyboard dialogues 86 out of 156 referential acts were used to refer to objects in the current focus of attention (55%). The 70 remaining referential acts (45%) were used to refer to objects outside of the current focus area (see chapter 3; Cremers and Beun (1995) for the criteria used to make this bipartition). Hence, no clear preference for choosing the next object in or out of the current focus area could be detected, as was the case in the spoken dialogues (68% in focus, 32% out of focus). This result confirms the expectation and is probably due to a coordination problem between typing and inspecting the domain.

Among the 86 references used in the keyboard dialogues to refer to objects within the current focus of attention, focus reduction was applied in 20 cases (23% of 86). This percentage is very close to that found in spoken dialogues, where focus reduction was applied in 27% of the cases.

Our prediction was, however, that in keyboard dialogues more cases of focus reduction would occur owing to a general reduction of words. The result seems to suggest that this was not the case. However, if we again consider the use of gestures as a means to reduce effort, some evidence for the truth of the hypothesis can be found.

Participants in keyboard dialogues used gestures without any language to refer to objects in 35 (41%) of the in-focus cases. In spoken dialogues this was done in 13 cases (13%), where the gesture was accompanied by just a demonstrative.

If we add the cases of gesture-related focus reduction to those where only a verbal reduction took place, the total number of cases of focus reduction in keyboard dialogues becomes 55 (64% of the in-focus cases). In spoken dialogues the total number of focus reduction then becomes 40 (40% of the in-focus cases). This suggests that, in the latter interpretation of focus reduction, participants in keyboard dialogues indeed use more reduced information when referring to objects within the focus area than do participants in spoken dialogues. However, this reduction is due more to the use of gestures than to the use of reduced verbal information. An overview of the findings is given in Table 4

Table 4 Focus reductions in keyboard and spoken dialogues.

| | **keyboard (86)** | **spoken (99)** |
|---|---|---|
| **verbal** | 20 (23%) | 27 (27%) |
| **gestures** | 35 (41%) | 13 (13%) |
| **total** | 55 (64%) | 40 (40%) |

## 5.4.2 Results concerning the process of object reference

### 5.4.2.1 Number of turns

In keyboard as well as spoken dialogues the mean number of both verbal and non-verbal turns needed to arrive at the mutual agreement that the target object has been identified is exactly the same, namely 2.7 (s.d.=1.04 and 1.38, respectively). However, this does not mean that the process is exactly the same for both types of dialogues. The difference lies in the relative use of verbal turns and (referential) actions in this process. In keyboard dialogues 98 (63%) of the turns were non-verbal, whereas in spoken dialogues gestures or actions were used only in 23 (16%) of the turns. No indication was found that more information was given in the first turn to avoid turn-takings since the mean lengths of first referential acts in keyboard and spoken dialogues were very similar (keyboard: 1.8, spoken: 2.2) and even shorter in keyboard dialogues.

With respect to the number of turns necessary to refer to objects in or out of focus a difference between spoken dialogues and keyboard dialogues was found. In spoken dialogues more turns were needed to refer to an object out of focus (3.2) than to one in focus (2.4), whereas no difference could be found in keyboard dialogues (both 2.7). This confirms our expectation that participants in keyboard dialogues do not benefit very much from the focus area, probably due to coordination problems between typing and inspecting the domain.

### 5.4.2.2 Miscommunications

One of the expectations presented in section 5.3.2.2 was that in keyboard dialogues more turns due to miscommunications would occur, since participants use fewer words to refer to objects. In the preceding section it was shown that no difference in the mean number of turns between keyboard dialogues and spoken dialogues occurred. This means that, if more miscommunications occurred, they did not increase the mean number of turns significantly. The results of analyzing the occurring miscommunications are given in Table 5.

Table 5 Miscommunications in keyboard and spoken dialogues.

| | keyboard (156) | | spoken (145) | |
|---|---|---|---|---|
| **focus** | 13 | (52%) | 5 | (83%) |
| **mistake** | 4 | (16%) | 1 | (17%) |
| **determiner** | 8 | (32%) | - | - |
| **focus (-determiner)** | 13 | (77%) | 5 | (83%) |
| **total** | 25 | (16%) | 6 | (4%) |

In keyboard dialogues miscommunications occurred in 25 (16%) of the cases before identification took place. In spoken dialogues only six (4%) of the first references to objects in the domain were initially misunderstood. These miscommunications were found to be due mainly to misunderstandings related to focus (in five cases, 83%). The one remaining case (17%) was due to a mistake made by the speaker.

In the keyboard dialogues 13 (52%) of the misunderstandings were in some way related to focus. In four cases (16%) mistakes were made by either one of the participants. In the remaining eight cases (32%) the misunderstanding was a result of confusion as to whether a new object should be introduced or the referential act was meant to refer to an object in the domain. These confusions were directly related to the fact that the typists did not add any determiner to the referential expression. This is a clear consequence of the modality of communication that was used. In order to type as few words as possible, typists omitted determiners thereby leading to a misunderstanding.

Since the latter group of misunderstandings was a direct result of the available modalities of communication, they can be omitted from the comparison between keyboard and spoken dialogues. The percentage of misunderstandings due to focus then becomes 77% (13 out of 17 cases), which is close to the 83% found in spoken dialogues.

To summarize, more or less the same percentage of focus-related misunderstandings occurred in keyboard dialogues as in spoken dialogues. However, the total percentage of misunderstandings in keyboard dialogues was greater since more misunderstandings occurred due to mistakes and, most importantly, due to omitting the determiner in the description. This result stresses the importance of determiners that provide information about the accessibility of the referent (see Piwek, Beun and Cremers, 1995; 1996).

## 5.5 Discussion and conclusions

The differences between the uses of referential expressions and gestures in keyboard and spoken dialogues can be explained to a large extent by the differences in the respective experimental paradigms as illustrated by the DenK triangle.

A direct consequence of the change from spoken to typed communication are the lengths of the referential expressions used. Since it takes more effort to type than to speak, fewer words were used in referential expressions in keyboard dialogues than in spoken dialogues. However, since the difference was not very great, the largest reduction of words occurred in the non-referential parts of the utterances. Furthermore, it could not be demonstrated that participants in keyboard dialogues used fewer gestures than those in spoken dialogues. The total number of gestures was the same although the distribution over accompanying features was different. However, these results may be domain-dependent since objects that are more difficult to describe are expected to be pointed at more often.

The difference in the distribution of gestures was a direct consequence of the problematic coordination of verbal and non-verbal information in keyboard dialogues. Since it was not possible to gesture and type at the same time, hardly any occurrences of short referential expressions, such as demonstratives or absolute features only, were found. In spoken dialogues the demonstratives and absolute features that accompanied gestures can be said to have the function of either attracting the attention of the partner to look at the domain or keeping the conversation flowing by avoiding silences. In keyboard dialogues the latter function is not very prevalent since the time pressure is not so great there (see Beun and Bunt, 1987). Participants in keyboard dialogues lost the possibility to apply the former function, i.e. to attract attention. However, these participants were observed to point with more emphasis, i.e. repeatedly or for a longer period than participants in spoken dialogues did. This emphasis can be interpreted as a means to make sure that the partner has observed the gesture.

A second consequence related to the coordination of modalities was the fact that typing and simultaneously inspecting the domain was difficult. This resulted in difficulty in keeping track of the current focus area. This difficulty was reflected in the same number of references to objects in focus to objects out of focus, compared to this distribution in spoken dialogues where far more references to objects in focus occurred.

As a consequence of the difficulty in changing turns in keyboard dialogues fewer verbal turns took place. However, the loss of verbal turns was compensated by more non-verbal turns. There was no indication that more information was given in the first utterance to try to avoid having to use more turns. However, this could be a consequence of the relatively simple objects used in the experiment. It was probably not necessary to use more words to indicate a certain object unambiguously. Although more miscommunications occurred in an absolute sense, they did not affect the mean number of turns used to reach common agreement that the target object had been identified.

The differences between keyboard and spoken dialogues were all found to be based on the principle of minimal cooperative total effort. In a situation where different modalities of communication are available which have different characteristics and possibilities, other means have to be found to minimize effort. The main change with respect to spoken dialogues was in the use of gestures to refer to objects. In both spoken dialogues and keyboard dialogues the same numbers of gestures were used, although they were used at different moments. At moments where participants in spoken dialogues used limited information, participants in keyboard dialogues tended to use more pointing gestures.

From these findings some implications can be drawn for the design of a multimodal interface, such as the DenK interface. First, in our domain we did not find a large reduction of words in referential expressions, but we did find a large reduction in the rest of the utterances, i.e. in the part were the action that has to be carried out is formulated. Further research should be conducted to figure out whether this reduction causes more or other types of miscommunications.

In the design of a multimodal interface special attention should be devoted to the coordination of verbal and non-verbal information. Procedures should be developed to make links between verbal expressions, especially longer ones, and gestures that are meant to refer to the same objects but do not occur at the same time. This is necessary in order to avoid confusions about whether in these cases only one object or two separate ones are being referred to.

In keyboard dialogues participants apparently did not make use of the current focus area as often as participants in the spoken dialogues, but reduced expressions referring to objects within the current focus area still occurred regularly. This means that the interface should adopt a notion of focus area in order to enable these expressions to be understood.

Finally, the interface should allow users to change turns quickly since almost the only type of feedback that was provided in the keyboard dialogues consisted of gestures or actions in the domain. It is probably easier for the interface to understand verbal feedback than to have to analyze the meaning of the gestures and actions. However, provisions should be made for listing the verbal and non-verbal turns in a convenient way so that no confusions will arise because the correct order of the turns is unclear.

# 6 Object reference and spatial focus: an experimental study

## *Abstract*

*This chapter describes an experimental study that was carried out to test the assumption that in a cooperative task-oriented dialogue the spatial focus of attention guides the addressee in choosing the target object on the basis of a certain referring expression uttered by the speaker. The spatial focus of attention is defined as the area within the task domain closely surrounding an object referred to previously.*

*Referring expressions can have three different grades of specification: ambiguous, minimally specified and redundant. This results in six possible combinations of grade of specification within in the focus area and within the entire domain: 1. ambiguous/ambiguous, 2. minimally specified/ambiguous, 3. minimally specified/minimally specified, 4. redundant/ambiguous, 5. redundant/minimally specified, and 6. redundant/redundant. It was hypothesized that, if a focus area is present, a descending preference from combination 1 to combination 6 for choosing the target object within the focus area would occur.*

*In the experiment a series of images displayed on a computer monitor was presented to the subjects. In each image six blocks were depicted, divided into two groups of three blocks and accompanied by a referring expression. The subjects were instructed to touch the block they thought was being referred to in the expression. Two conditions were tested: a focus condition for which a spatial focus area had been established beforehand, and a no-focus condition for which this area was not established.*

*The results show a significant overall influence of the spatial focus area on the choice of the target object. In addition, the preference for choosing an object in the focus area gradually becomes smaller from combination 1 to 6, and eventually turns into a preference for choosing an object in the non-focus area.*

*The results provide some evidence that participants in task-oriented dialogues take into account the spatial focus area when generating and interpreting referring expressions. However, in a real dialogue situation, in which the referring expression is usually embedded in a larger utterance, other factors, such as functional focus, play a role as well.*

## 6.1 Introduction

When performing goal-directed behavior, humans generally try to use as little effort as possible to achieve their goals. In a situation where two humans are carrying out a task cooperatively, they can strive for a minimization of effort, among other things, in the process of generating and interpreting contributions to the associated dialogue. This minimization of effort also applies to the use and understanding of referring expressions and/ or gestures to refer to objects that are involved in the task.

There are at least two ways in which the latter type of minimization of effort can be established cooperatively by the speaker and the hearer. First, the speaker may try to reduce effort by using absolute features in the referring expression (as opposed to relative features) and by limiting the number of words used as much as possible. Using absolute features means less effort for both the speaker and the hearer, since the target object does not have to compared to other objects. Fewer words mean less effort to generate as well as interpret the utterance.

Second, the speaker and the hearer may minimize effort by trying to reduce the number of objects to be considered as possible target objects. The advantage for the speaker is that, in general, fewer features have to be included in the referring expression to distinguish the target object from the competing objects. This is also an advantage for the hearer, who has to consider fewer objects in order to find the target object. An important way to reduce the number of objects to be considered as potential target objects is to take the current *spatial focus of attention* into account. The spatial focus of attention covers an area, called the *focus area*, that is for some reason in the current focus of attention of the participants, for instance, because the area surrounds the previous object that has been referred to.

Assuming that the spatial focus of attention is taken into account by both the speaker and the hearer, it is sufficient for the speaker to include just the type and/or amount of information in the referring expression to allow the hearer to distinguish the target object from other objects within the focus area. In many situations, such a minimally specified expression would be ambiguous if the entire domain of conversation were to be considered, especially in domains where more than one object of the same type occurs. In comparison with ambiguous expressions (that do not provide enough information to distinguish the target object from other objects within the focus area) and redundant expressions (that provide more information than is strictly necessary to distinguish the target object from other objects within the focus area), minimally specified expressions seem to cost the least amount of cooperative total effort.

The main goal of this chapter is to report on an experimental study that was carried out to test the assumption that the spatial focus of attention guides the addressee in choosing the target object on the basis of a certain referring expression. In particular, it was investigated how the grade of information included in the referring expression affects this expected influence.

In section 6.2 the theoretical background with respect to the concepts of focus of attention and redundancy of referring expressions is outlined in the framework of the principle of minimal cooperative total effort. In section 6.3 the main hypotheses are formulated. In section 6.4 the experimental method is presented that was used to test these hypotheses, including a detailed description of the conditions and the types of referring expressions that are used as stimuli. The results of the experiment are presented in section 6.5 and discussed in section 6.6. Finally, in section 6.7 some conclusions are drawn and future research is suggested.

## 6.2 Theories on reference, minimal effort and focus of attention

### 6.2.1 Reference and minimal effort

If a speaker wants to refer to some object located in a domain that can be perceived by the hearer, both participants have to get through several steps. The speaker has to decide first which object to refer to. Next he or she has to formulate the referring expression. The hearer has to interpret this expression and, finally, identify an object that he or she considers to be the intended target object. This process may take several dialogue turns and wrong identifications before the dialogue participants arrive at mutual agreement that the target object has been identified.

The choice to use a certain referential expression in a natural human-human dialogue is based on a number of considerations on the part of the speaker. In particular, the choice of the amount of information, and the features of the target object to be described in the expression are based on considerations with respect to the effort it takes to utter this expression. This view of object reference has a long tradition. With respect to the amount of information it is assumed that effort can be reduced by using as few words as possible. Grice formulated this in his *maxim of quantity* (Grice, 1975), which says: (1) make your contribution as informative as is required for the current purposes of the exchange, and (2) do not make your contribution more informative than is required.

The same idea was formulated more specifically in terms of object reference in the *principle of minimal effort* (Brown, 1958; Krauss & Glucksberg, 1977; Olson, 1970). According to this principle speakers try to utter a noun phrase that is as short as possible, while still allowing the hearer to select the intended referent. More recently Clark and Wilkes-Gibbs (1986) have taken the view that this minimal effort is of a cooperative nature, which means that dialogue partners have a joint responsibility to use as few words as possible to reach the conclusion that the target object has been identified. Therefore, they rechristened the principle as the *principle of minimal cooperative effort*. In other words, participants are aware of a trade-off between the number of words in the initial noun phrase and the need for the speaker to refashion this noun phrase, or for the hearer to ask for clarification.

However, the total amount of cooperative effort that is used to identify an object is not only of a linguistic nature. Therefore Cremers and Beun (1995; chapter 3) extended Clark and Wilkes-Gibbs's (1986) principle to the *principle of minimal cooperative total effort*, hereby indicating that besides the linguistic effort, the effort to actually identify the target object should also be taken into account. This effort is also of a cooperative nature; the speaker can, for instance, decrease the effort of the hearer to identify the object by using features of the target object that are salient within the domain of conversation.

### 6.2.2 Reference and focus of attention

An important means to reduce cooperative effort is to make use of the current focus of attention. Entities that have been mentioned very recently in a dialogue are in focus and can usually be referred to by means of very few words, such as a pronominal expression. Entities that have not been mentioned explicitly but that are associated with an entity in focus can be said to be in implicit focus, and can accordingly be referred to by means of a definite expression (e.g., if the entity referred to by 'the book' is in focus, the associated entity referred to by 'the author' is in implicit focus).

Grosz and Sidner (1986) showed that in task-oriented dialogues concerning the installation of a water pump, the task structure determines which objects are in explicit or implicit focus. For instance, the definite expression 'the screw' can be used to refer to the particular screw that is part of the component that is being installed at that moment. Because of the existence of the task structure this expression does not cause an ambiguity with respect to screws present in other parts of the pump.

As one of the results of an empirical study, Cremers and Beun (1995; chapter 3) found that an important means to minimize the linguistic *and* the non-linguistic effort in spoken first references to objects in a spatial domain of conversation was for the participants in a task-oriented dialogue to make use of the *spatial focus of attention.*[1] The task was for one participant to instruct the partner to make some changes in a block building on the basis of an example building that was not visible to the other. An object could become part of the current spatial focus of attention in this particular domain if it was located close to the object that had been mentioned previously (in other words, if it was part of the current *focus area*).

In 27% of the references to objects within the current focus area speakers used exactly enough information to distinguish the target object from the other objects within this area, but not enough to distinguish the target object from all objects present within the whole domain. This behavior can be interpreted as an attempt to minimize effort for both speaker and hearer. Based on the assumption that the hearer is focusing attention at the focus area, the speaker only has to include a minimal amount of information in the referring expression. If the assumption is right then the hearer has to consider fewer objects in order to find the target object (only the objects within the focus area).

At focus transitions, i.e., at places during the interaction where a new area of the block building became subject to alterations, speakers had to expend more effort than at places where the focus area remained the same, to make clear that a new focus area had to be established. At these points they either used explicit focus transition indicators (e.g., 'let's go the upper left-most part now') or they included sufficient information in

[1] Although this result was replicated in an empirical study on keyboard dialogues (Cremers, 1995b; chapter 5), in this chapter reference is only made to the results of the study on spoken dialogues.

the referential expression itself to distinguish the target object from all other objects present in the domain (e.g., 'the block lying horizontally, the blue bar of two by six, in the front').

### 6.2.3 Grades of specification in references

Basically, a referring expression can have three different grades of specification: *ambiguous, minimally specified* or *redundant*. In terms of effort, using an ambiguous expression means less initial effort for the speaker, but more effort thereafter for the hearer to attempt identification and to ask for clarification, and again by the speaker to expand or to refashion the initial expression. Minimal specification seems to result in the least amount of cooperative effort for both the speaker and the hearer, since exactly enough information is provided to allow the hearer to identify the target object. Redundant expressions seem to cost too much effort, in particular for the speaker, since he or she provides more information than is strictly necessary. The hearer may benefit from this redundancy, resulting in a reduction of effort expended to identify the referent.

Although minimal specification appears to be the most efficient way to refer for both speaker and hearer, this does not seem to be the prototypical referential behavior. There is some experimental evidence that speakers have a tendency to *overspecify* in referring expressions, i.e., give redundant information (Deutsch and Pechmann (1982): in 28% of the references; Pechmann (1984): in 60% of the references).

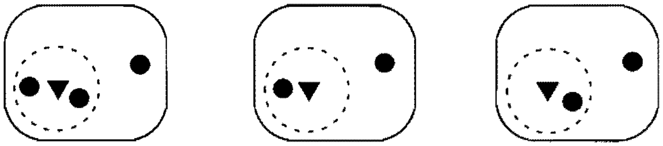
One could argue that a certain amount of redundancy is inherent to language use, but that does not exclude the possibility that speakers may have a reason for using redundant information. Levelt (1989) lists two reasons why speakers overspecify in referring expressions. In the first place, redundant information can help the addressee to find the referent. For example, if the *type* of object is mentioned but this information is redundant in the context of the discourse, it still helps the addressee to create a *Gestalt* of the object which makes it easier to find it. Such a situation occurs for instance if some pyramids and some other objects are present within the domain of conversation, but only one of the objects, which happens to be a pyramid, is yellow, and this block is referred to by means of 'the yellow pyramid'. If the addressee then creates a Gestalt of a pyramid, it helps him finding the only yellow object. In the second place, redundant information can be provided to contrast the object with the last one focused on by the listener. This can be done if the current referent only differs slightly from the previous one, for example in 'the yellow pyramid, the red pyramid', where 'red' is given a contrastive accent.

In the spatial blocks domain of the empirical study by Cremers and Beun (1995; chapter 3), it was not necessary to create a new Gestalt of every target object, since almost all objects involved were blocks with similar shapes and only a limited number of sizes. Most of the referring expressions that were used did not include a Gestalt-inducing element. However, the second reason for overspecification given by Levelt (1989), i.e., contrast, is valid in the blocks domain. Since many blocks were similar in shape and color there was a high probability that the current referent would differ only slightly from the previous one (e.g., 'the large red one, the small red one'). A specific feature of the blocks domain is that blocks can also be located close to one another spatially, and that this proximity can be used for contrasting two referents, e.g., 'the red block next to the blue one' where 'the blue one' has been mentioned previously.

### 6.2.4 Grades of specification and focus

Given that redundancy is apparently common in referring expressions and that it helps the addressee to find the referent, it is the more remarkable that expressions that were ambiguous if the entire domain was taken into account, occurred so often in Cremers and Beun's (1995; chapter 3) dialogues. However, if we assume that the participants only considered the spatial focus area, these same expressions were minimally specifying within this area, i.e., they provided exactly the amount of information that results in minimal effort. Therefore the spatial focus of attention turned out to be a very strong device to help the hearer to interpret ambiguous referring expressions.

In a spatial domain the different grades of specification can be formulated in a more specific way. Since a distinction can be made between blocks that are located within the current focus area and those that are located outside, the grade of specification of a referring expression can be determined either with respect to the objects in the focus area or those in the entire domain. If we consider the situation where at least two identical objects are located within the domain, and at least one of these objects is located within the focus area and at least one outside, in total six combinations of grades of specification can occur. If the expression is ambiguous in the focus area it is also ambiguous in the whole domain (see Figure 14a). If the expression is minimally specified in the focus area it can be either ambiguous (see Figure 14b) or minimally specified in the whole domain. Finally, if the expression is redundant in the focus area it can be either ambiguous, minimally specified (see Figure 14c), or redundant in the whole domain.



a. 'the circle' b. 'the circle' c. 'the circle on the right'

Figure 14 Examples of references to objects in spatial domains (The area indicated by the dashed line is the current focus area.): a. ambiguous in focus area and in whole domain; b. minimally specified in focus area, ambiguous in whole domain; c. redundant in focus area, minimally specified in whole domain.

The results of the empirical study by Cremers and Beun (1995; chapter 3) suggest that in a situation where the referring expression is minimally specified within the focus area, the object within the focus area that meets the description will be chosen by the hearer. In the case of an expression that is ambiguous within the focus area, it is expected that hearers will also search for the target object in the focus area, and then realize that not enough information is provided to find the intended target object. When an expression is used that is redundant if only the focus area is considered, it is not obvious what the influence of the focus area will be. Given the empirical evidence that a speaker tends to use minimally specified information to refer to an object in the focus area and more elaborate information if a focus transition is taking place, it may be argued that the hearer will choose a suitable object outside of the focus area as the intended target object.

## 6.3 Hypotheses

The considerations in the previous section lead to two main hypotheses concerning the identification of a certain object in a spatial domain on the basis of a referring expression with a certain grade of specification, depending on whether a certain focus area is present or not.

**Hypothesis 1 (Focus)**

- *If* no focus area is present in a spatial domain (for instance, because no earlier references to objects in this domain have taken place),
  *and* the speaker uses a referring expression that has at least two possible referents within the domain,
  *and* the hearer is forced to identify an object on the basis of this expression,
  *then* the chance that one particular object of the set of objects that meet the description is chosen is the inverse of the number of elements in the set.
- *If* a focus area is present in a spatial domain (for instance, because a certain object has been referred to just before by the speaker and has been identified by the hearer),
  *and* the speaker uses a referring expression that has at least two possible referents within this domain,
  *and* at least one object meeting the description is present within the focus area and at least one is present outside of it,
  *and* the hearer is forced to identify an object on the basis of this expression,
  *then* the chance that one particular object of the set of objects that meet the description is chosen is influenced by the existence of the focus area, in the sense that a preference for choosing an object either within the focus area or outside of the focus area will occur. This preference depends on the grade of specification of the referring expression (see Hypothesis 2).

Three grades of specification can occur in referring expressions: *ambiguous*, *minimally specified* and *redundant*. In a spatial domain these grades of specification can apply to either just the focus area or the entire domain, resulting in six different combinations: 1. ambiguous (within the focus area) vs. ambiguous (within the entire domain), 2. minimally specified vs. ambiguous, 3. minimally specified vs. minimally specified, 4. redundant vs. ambiguous, 5. redundant vs. minimally specified, and, 6. redundant vs. redundant. The hypothesis concerning the grade of specification is that:

**Hypothesis 2 (Grade of specification)**

- *If* a focus area is present in a spatial domain,
  *and* the speaker uses a referring expression that has at least two possible referents in this domain,
  *and* at least one object meeting the description is present within the focus area and at least one is present outside of it,
  *and* the hearer is forced to identify an object on the basis of this expression,
  *and* the expression is an instance of combination 1. of grades of specification,
  *then* the hearer prefers to choose an object within the focus area.
- The preference for choosing an object within the focus area will gradually become weaker as the combination of grades of specification changes from 1. to 6. consecutively, and will eventually even turn into an increasing preference for choosing an object in the non-focus area.

## 6.4 Experimental method

An experiment was designed to test these hypotheses. In the following the experimental design is described in detail, with special emphasis on the stimuli that were used.

### 6.4.1 Subjects

In the experiment 20 Dutch subjects participated (10 males and 10 females). Their ages varied from approximately 25 to approximately 45.

### 6.4.2 Equipment

A Macintosh® was used to run the experiment. The experiment was programmed in Psyscope® (Cohen et al., 1993). A touch screen was used as input device.

### 6.4.3 Material

As experimental material 252 trials were designed that included 'blocks'[2] domains accompanied by referring expressions. The blocks domains consisted of six blocks in different combinations of the colors red, blue, green, yellow, grey and black. The domains were designed according to the experimental conditions that were to be tested (i.e., presence of focus and grade of specification of referring expressions) and will be described in more detail later. Of the 252 trials, 216 were experimental trials and 36 were fillers. Fillers were used to make sure that the subjects did not become aware of the underlying experimental question. In the filler trials there was never a choice problem; the expressions used in these trials referred unambiguously to one of the blocks in the domain.

The six blocks were all equally sized 2.5 by 2.5 cm, they were placed in a horizontal line and divided into two groups separated from each other by a space exactly the size of one block (see Figure 15). This arrangement was chosen to try to ensure that the sub-

[2] Actually, the domains were two-dimensional, so squares rather than blocks were depicted. They are still called blocks, as was done during the experiment.

jects would perceive either one of the two groups as the current focus area. Colors were chosen randomly for each blocks domain from the six colors available to fill the blocks according to the experimental conditions. The referring expressions (in Dutch) were located 3 cm beneath the blocks, in 18-point character size, in lower case, and centered on the screen.

Three different types of features were used in the expressions, since these were all considered to be possible distinguishing features within this particular domain. They were: color (e.g., 'grey'), absolute location (e.g., 'right'), relative location (e.g., 'next to') and combinations of these.

### 6.4.4 Conditions and groups

To test the effect of the spatial focus of attention on the identification of objects, all 252 trials were presented twice during the experiment: once in the *no-focus condition* (where it was ensured that no focus area was present), and once in the *focus condition* (where it was assumed that a focus area was present). In the focus condition the focus area, which was assumed to be one of the two groups of three blocks, was established by a preceding expression unambiguously referring to one of the blocks in this area. In the no-focus condition no preceding referring expression occurred.

To test the influence of the *six different combinations of grades of specification*, the 216 non-filler trials were built up out of six groups of 36 trials that each represented one of these combinations (see Table 6 for examples). In each group all possible combinations of focus-establishing blocks and target blocks occurred. There were as many focus-establishing blocks located in the left area as in the right area, to ensure that the focus area was in the left area in half of the trials and in the right area in the other half.

The referring expressions that were used always had at least two possible referents: one within the focus area and one outside. In most of the groups it was not possible to design 36 trials on the basis of just one type of referring expression. In these groups more than one type of expression was used, which is indicated by the subgroups.

A more detailed description of the six groups of stimuli that correspond to the different combinations of grades of specification is given in the following subsections, based on the examples in Table 6 When necessary, more specific hypotheses dealing with particular details of the stimuli, in particular in the case of subgroups, are presented here as well.

#### 6.4.4.1 Group 1

In group 1, the expression (of the type 'de grijze' ('the grey one')) is not only ambiguous when just the focus area is taken into account, but also when the whole domain is considered. The focus area always contains two identical blocks (in the example, two grey blocks). In subgroup a. one of these blocks is also present in the non-focus area, and in subgroup b. two of these blocks are present there.

In the no-focus condition it is expected that subjects will choose any one of the three or four potential target objects with equal likelihood. In the focus condition, a preference for choosing one of the possible referents in the focus area is expected. Instead, in subgroup a., a preference for the non-focus area may occur, due to the uniqueness of the possible referent within this area.

#### 6.4.4.2 Group 2

In group 2, the referring expression (of the type 'de grijze' ('the grey one')) is ambiguous within the whole domain, but it is minimally specified within the focus area. In the example there are two grey blocks within the whole domain, but only one within the focus area. In the no-focus condition it is expected that subjects will choose evenly between the two potential target objects. In the focus condition a clear preference for choosing the object within the focus area is expected, even more so because the information is ambiguous within the whole domain.

#### 6.4.4.3 Group 3

In group 3, minimally specified information is provided to find the target object within either the whole domain or the focus area. In subgroup a. the expressions 'de rechtse/linkse' ('the right/left one') were used. In subgroup b. the expressions 'de meest rechtse/linkse' ('the right-/left-most one') were used. In the subgroups c. and d. these same expressions were used, but here the focus-establishing block was the same as either the right or the left block within the focus area.

In the no-focus condition a preference for choosing the block that is the absolute right/left one in the whole domain is expected. In the focus condition more preference for choosing the object in the focus area is expected, i.e., the right-/left-most block within the focus area. In subgroup a. the predicate 'right/left' may overrule the preference for the focus area if it is taken to mean the absolute right/left block within the whole domain. In this case no significant difference with the no-focus condition will be found. In subgroup b. this effect may be even larger than in subgroup a. In the subgroups c. and d. a reason for switching to the non-focus area may be to avoid having to choose the same object two times in a row.

#### 6.4.4.4 Group 4

In group 4, the information to find the referent within the focus area is redundant, but it is ambiguous for finding the referent within the whole focus area. In subgroup a. the two blocks mentioned in the referring expression (of the type 'de grijze naast de zwarte' ('the grey one next to the black one')) are always adjacent. In the example there is only one grey block present within the focus area, so the addition of 'next to the black one' is redundant here. In the whole domain there are two grey blocks adjacent to black ones, so the information is ambiguous there. In subgroup b. the focus-establishing block is located in between the two blocks mentioned in the referring expression (of the type 'de grijze rechts/links naast de zwarte' ('the grey one right/left of the black one')), whereas they are adjacent in the non-focus area.

In the no-focus condition it is expected that subjects will not have a preference for choosing either of the potential target objects in subgroup a. They may have a preference for choosing the target object that is adjacent to the relatum in subgroup b., since they may take the expression 'to the right/left of' to mean 'adjacent to the right/left of'. In the focus condition in subgroup a., a shift is expected to the non-focus area due to the redundancy of the information. The shift may be even larger in subgroup b., because of the adjacency of the blocks involved in the non-focus area.

#### 6.4.4.5 Group 5

Group 5 represents the situation where the information to find the object in the focus area is redundant, but is minimally specified in the whole domain. In subgroup a. expressions like 'de rechtse/linkse grijze' ('the right/left grey one') were used. In the example either 'the grey one' or 'the right one' would be sufficient within the focus area; however, since there are two grey blocks in the whole domain 'the right grey one' constitutes a minimal specification in this domain. In subgroup c., expressions like 'de meest rechtse/linkse grijze' ('the right-/left-most grey one') are used. In subgroups b. and d., where both of the above types of expressions were used, the focus-establishing block was in the same area as the block that would be chosen with respect to the whole domain. In subgroups e. and f. the focus-establishing block is even the same as the block that would be chosen with respect to the whole domain.

In the no-focus condition a preference for choosing the right-/left-most block within the whole domain is expected in all subgroups. A specific preference may appear in the cases where one of the possible target objects is placed at the right-/left-most position within the assumed 'focus area', and where the second target object is placed in the non-focus area, but not in the right-/left-most position. In these cases a preference for choosing the object within the 'focus area' may occur, since it is the only object in a right-/left-most position within a group of three blocks.

In the focus condition, a shift is expected towards the object in the non-focus area due to the redundancy of the expression within the focus area. Again, the predicates 'right/left' may overrule the still existing, possibly small, preference for the block within the focus area, which will result in an even larger preference for the non-focus area. For the subgroups b. and d. a preference for the focus area is expected, because the right-/left-most referent in the domain is the same as the one in the focus area. In the subgroups e. and f. it is not so clear what the response of the subjects will be. They will either stay in the focus area, or they will switch to the other area to avoid having to choose the same block twice.

Table 6 Groups of stimuli used in the experiment. (The grade of specification within the focus area and within the whole domain are given for each group, the number of stimuli and an example of a stimulus are indicated for each subgroup. The block indicated by '*' was referred to just before by means of a minimally specified expression, in order to establish a focus area. The white blocks are actually meant to be colored either blue, yellow, red or green (or grey or black in the examples in group 3.))

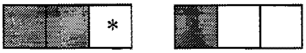
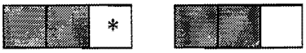
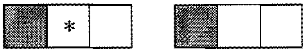
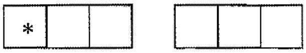
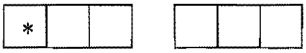
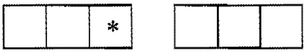
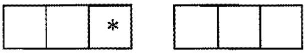
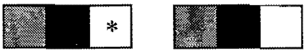
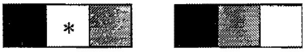
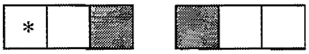
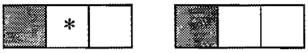
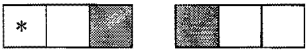
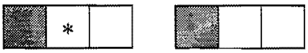
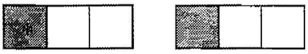
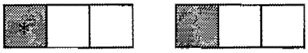
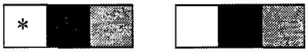
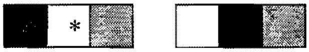
| group | grade spec. focus area/ domain | sub-group | number of stim. | examples |
|---|---|---|---|---|
| 1. | ambiguous/ ambiguous | a | 18 |  the grey one |
| | | b | 18 |  the grey one |
| 2. | minimally specified/ ambiguous | | 36 |  the grey one |
| 3. | minimally specified/ minimally specified | a | 12 |  the right one |
| | | b | 12 |  the right-most one |
| | | c | 6 |  the right one |
| | | d | 6 |  the right-most one |
| 4. | redundant/ ambiguous | a | 32 |  the grey one next to the black one |
| | | b | 4 |  the grey one to the right of the black one |

Table 6 (Continued) Groups of stimuli used in the experiment. (The grade of specification within the focus area and within the whole domain are given for each group, the number of stimuli and an example of a stimulus are indicated for each subgroup. The block indicated by '*' was referred to just before by means of a minimally specified expression, in order to establish a focus area. The white blocks are actually meant to be colored either blue, yellow, red or green (or grey or black in the examples in group 3.))

| group | grade spec. focus area/ domain | sub-group | number of stim. | examples |
|---|---|---|---|---|
| 5. | redundant/ minimally specified | a | 12 |  the right grey one |
| | | b | 4 |  the left grey one |
| | | c | 12 |  the right-most grey one |
| | | d | 4 |  the left-most grey one |
| | | e | 2 |  the left grey one |
| | | f | 2 |  the left-most grey one |
| 6. | redundant/ redundant | a | 24 |  the right grey one next to the black one |
| | | b | 12 |  the right grey one to the right of the black one |

### 6.4.4.6 Group 6

Finally, in group 6, the information both to find the target object in the current focus area and in the whole domain is redundant. In subgroup a. expressions such as 'de rechtse/ linkse grijze naast de zwarte' ('the right/left grey one next to the black one') are used. In the example clearly too much information is provided to identify the unique grey block within the focus area, and also when considering the whole domain the expression 'the right grey one' would have sufficed to identify the target object. In sub-condition b. expressions such as 'de rechtse/linkse grijze rechts/links van de zwarte' ('the right/left grey one to the right/left of the black one') were used.

In the no-focus condition in subgroup a., a preference for choosing the right-/left-most potential target object within the whole domain is expected to occur. Exceptions may be the case where the potential target object in the non-focus area is not placed in the right-/left-most position. In that case subjects may exhibit a preference for the potential target object within the 'focus area', which is always located right-/left-most within this area. In subgroup b. a preference for the non-focus area is also expected, since the two blocks mentioned are located adjacent to each other there.

In the focus condition it is expected that an even larger preference for choosing the target object in the non-focus area will occur than in the no-focus condition, due to the redundancy of the expression within the focus area.

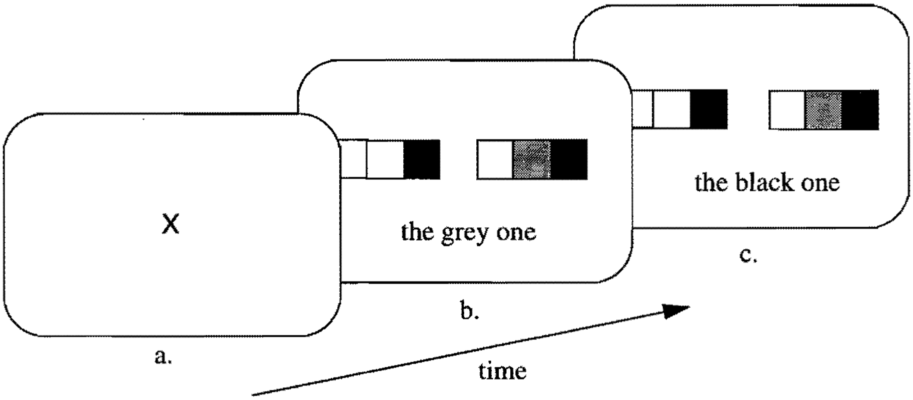


Figure 15 The consecutive images being displayed on the monitor (called 'screens' in short) used in the focus condition (a., b., c.) and in the no-focus condition (a., c.).

### 6.4.5 Procedure

The experiment consisted of two sessions, one for the no-focus condition and one for the focus condition. The order in which the sessions were presented to the subjects was balanced for the 20 subjects. In each session 252 trials (including 36 fillers) were presented to the subjects in random order. During the sessions four breaks were included that divided the trials in five blocks of about 50. The subjects were allowed to rest as long as they wanted here, but for at least ten seconds. Furthermore, the time interval of at least half a day between the two sessions was sufficiently long for the subjects to be able to rest.

At the start of a session the subjects were seated in front of the Macintosh monitor, such that the distance from their eyes to the monitor was approximately 40 centimeters. They were instructed that a series of blocks domains accompanied by referring expressions would appear on the monitor, and that they were supposed to touch the block that they thought was being referred to in the expression. They were told that in case of doubt (for instance, when the expression appeared to be ambiguous), they should choose the first block that appeared to be appropriate, as quickly as possible. After having practiced 5 trials in the presence of the experimenter, the 252 trials (including 36 fillers) were presented to the subjects.

A trial was built up differently for the two conditions, as is shown in the example depicted in Figure 15. In the focus condition all of the three screens were presented to the subject, whereas in the no-focus condition screen b. was omitted.

A trial always started with a fixation cross (screen a.) that was presented for 2 seconds, and that was located exactly in between the two groups of blocks that were to appear in the next stage.

Next, but only in the focus condition, a blocks domain appeared accompanied by a minimally specifying referring expression (screen b.). This screen was presented to establish the focus area. The task of the subject was to touch the block as quickly as possible that, according to him or her, was most probably the one being referred to in the expression.[3] Immediately after touching the screen the text disappeared. If the subject did not respond quickly enough, then the text disappeared after 5 seconds. This was done to make sure that the subject would follow his or her first intuition and not hesitate too long before choosing a particular block.

Subsequently a second referring expression appeared beneath the same blocks domain (screen c.). This referring expression contained information in one of the six different grades of specification. At this point there were always two or more potential referents, with at least one being in the focus area and at least one outside of the focus area. The subject again had to touch the block he or she thought the expression was referring to. After the subject's response both blocks domain and referring expression disappeared and a new trial started.

In the no-focus condition exactly the same 252 trials were presented, except that screen b. was omitted, so that the fixation cross was followed immediately by screen c. Hence, in this condition no focus area was established.

The recorded data were the coordinates of the locations on the screen that had been touched by the subject, as well as the response times of the subjects. These were the times measured starting from the appearances of the referring expression to the points at which the screen was touched.

## 6.5 Results

The data were transformed in the following way. If a subject had chosen the target object located within the focus area[4], the data point (i.e., the coordinates of the location on the

[3] In this context 'as quickly as possible' means that the subject should follow his or her first intuition and not hesitate too long before giving the response.

screen) was transformed into a one. If a subject had chosen the target object outside of the focus area or had chosen a wrong object, a zero was assigned to the data point.

The data were analyzed in two ways. First, an ANOVA was carried out with a 2 (focus) × 5 (number of groups) design with subjects as replications, to determine the overall interaction between the two conditions focus and no-focus with the groups containing different grades of specification of the referring expression. The ANOVA was used for analyzing the groups or subgroups 1, 2, 4a, 5a+c and 6a only. The other groups and subgroups were not included for reasons that will be discussed below where the results for the separate groups are provided. Second, separate analyses were carried out for each individual group and subgroup using t-tests (see Lewis (1993, p. 83-89) for a justification of this type of analysis).

The difference of the assigned value (one or zero) between the no-focus and the focus condition of the same trial was used as the dependent variable of the ANOVA. The main effects of both focus and group are significant: respectively F(1,19)=81.30, p<0.0001 and F(4,19)=526.43, p<0.0001. The interaction effect of focus × group is significant as well: F(4,19)=27.73, p<0.0001.

The results of the paired two-tailed t-tests that were used to analyze the individual groups are provided below and listed in Table 7 The t-values are the paired mean differences between the assigned values (one or zero) in the no-focus condition and in the focus condition. This means that if the t-value is positive, there was a greater preference for choosing the target object outside of the focus area in the focus condition than there was in the no-focus condition. If the t-value is negative, there was a greater preference for choosing the target object within the focus area in the focus condition than there was in the no-focus condition. In addition, the percentual differences between the occurrences of the target object within the focus area in the no-focus condition and in the focus condition are indicated. In the last two columns the numbers of missing values as well as examples of the referring expressions concerned are given.

### 6.5.1 Group 1

In group 1, 25.8% more target objects were chosen within the focus area in the focus condition than were chosen in the corresponding area in the no-focus condition (t=-11.801, p<0.0001). In subgroup a., the perceptual difference is 21.4% (t=-6.89, p<0.0001), and in subgroup b. it is 30.2% (t=-9.865, p<0.0001).

In subgroup a., a higher percentage of the target objects was chosen within the focus area in the no-focus condition than was chosen in subgroup b. (respectively 57.5% and 48.3%). At first this result seems to contradict the hypothesis about the preference for the non-focus area in subgroup a., due to the uniqueness of the potential target object within this area. However, if we look at the individual potential target blocks, we see that in subgroup a. the chance that one particular target block was chosen in the focus area is 57.5%/2=28.8%, whereas the one potential target object in the non-focus area was cho-

[4] Although in reality of course no focus area existed in the no-focus condition, it was taken to be the same area as in the one in the corresponding trial in the focus condition.

sen in 42.5% of the cases, which yields a preference for the non-focus area of 13.7%. In subgroup b. the percentages for the focus area and the non-focus area are almost the same (respectively 24.2% and 25.9%, which yields a difference of 1.7%).

### 6.5.2 Group 2

In group 2, a preference for the target object within the focus area was again found in the focus condition, whereas there was no preference for either area in the no-focus condition (t=-8.019, p<0.0001). However, the difference (18.2%) is somewhat smaller than in group 1.

### 6.5.3 Group 3

In group 3, no difference could be found between the two conditions. Apparently expressions including 'right' and 'left' were always taken by the subjects to mean the 'right-most/left-most' *within the whole domain.* In all subgroups the same strategy was followed. Since no effect of the focus area was found at all in this group due to the apparent strength of the words 'right' and 'left', the group was excluded from further analysis.

### 6.5.4 Group 4

In group 4, a small preference for the non-focus area was found in the no-focus condition (41% of the potential target objects were chosen within the focus area). This preference is due to subgroup b., where a large preference for the non-focus area occurs due to the adjacency of the blocks mentioned in the expression that were located there (only 2.5% of the potential target objects were chosen within the focus area). In subgroup a. no preference for either area was found in the no-focus condition (45% of the potential target objects were chosen within the focus area).

In the focus condition a significant preference was found for choosing the target object within the focus area, although this preference is smaller than in group 2 (t=-4.802, p<0.0001, difference: 10.9%). In subgroup a. the preference is about the same as in the whole group (t=-4.829, p<0.0001, difference: 12%). In subgroup b. the subjects hardly ever chose for the target object within the focus area (in 2.6% of the cases), which was about the same behavior as in the no-focus condition.

Since the effect of the adjacency of the blocks involved in subgroup b. clearly overrules the effect of the focus area, and since the number of trials in this subgroup was low anyway (4 trials per condition for each subject), this subgroup was excluded from further analysis.

Table 7 Results of the t-test for the no-focus/focus conditions per group and subgroup; percentual differences between the choice for the target object in the focus area in the no-focus and the focus conditions; missing values; examples of referring expressions per group.

| group | sub-group | paired t-value | probability (2-tailed) | difference -/+ focus (%) | miss. val. | example |
|---|---|---|---|---|---|---|
| 1 | (all) | -11.801 | <0.0001 | 52.9-78.7 = -25.8 | 1 | the grey one |
| | a | -6.890 | <0.0001 | 57.5-78.9 = -21.4 | - | |
| | b | -9.865 | <0.0001 | 48.3-78.5 = -30.2 | 1 | |
| 2 | (all) | -8.019 | <0.0001 | 50-68.2 = -18.2 | 3 | the grey one |
| 3 | (all) | 0 | <1 | 0.4-0.4 = 0 | 1 | the right(most) one |
| | a | 0 | <1 | 0.8-0.8 = 0 | - | |
| | b | - | - | 0-0 = 0 | 1 | |
| | c | 1 | <0.3193 | 0.8-0 = 0.8 | - | |
| | d | -1 | <0.3193 | 0-0.8 = -0.8 | - | |
| 4 | (all) | -4.802 | <0.0001 | 41-51.9 = -10.9 | 5 | the grey one next to/ to the right of the black one |
| | a | -4.829 | <0.0001 | 45.8-57.8 = -12.0 | 2 | |
| | b | 0 | <1 | 2.5-2.6 = -0.1 | 3 | |
| 5 | (all) | 1.772 | <0.0768 | 36-34.3 = 1.7 | - | the right(most)/ left(most) grey one |
| | a | 0.727 | <0.4680 | 5.4-4.2 = 1.2 | - | |
| | b | 0.815 | <0.4176 | 97.5-95 = 2.5 | - | |
| | c | 2.528 | <0.0121 | 5.4-1.3 = 4.2 | - | |
| | d | -0.445 | <0.6576 | 96.3-97.5 = -1.2 | - | |
| | e | -1 | <0.3235 | 97.5-100 = -2.5 | - | |
| | f | -1 | <0.3235 | 97.5-100 = -2.5 | - | |
| 6 | (all) | 2.438 | <0.0150 | 8.9-5.8 = 3.1 | 17 | the right grey one next to/to the right of the black one |
| | a | 2.336 | <0.0199 | 12.2-8 = 4.2 | 7 | |
| | b | 0.706 | <0.4807 | 2.1-1.3 = 0.8 | 10 | |

### 6.5.5 Group 5

In group 5, a preference for choosing the object outside of the focus area was found in the no-focus condition (36% of the target objects were chosen within the focus area). In the focus condition the choice for the focus area is somewhat smaller (34.3%), but the difference is not significant (t=1.772, p<0.0768).

Of the six subgroups only subgroup c. yields a significant difference between the no-focus and the focus conditions (t=2.528, p<0.0121). In both conditions, a large preference for choosing a target object outside of the focus area occurs (respectively only 5.4% and 1.3% of the target objects were chosen within the focus area). This is the area where the right-/left-most target object was always located. Subgroup a. is almost identical to subgroup c.; there is only a difference with respect to the expressions used ('right/ left' versus 'right-/left-most'). However, in this subgroup no significant differences between focus and no focus were found (t=0.727, p<0.4680). If we take the two conditions together the difference is still significant (t=2.273, p<0.0235, difference: 2.7%).

In the subgroups b., d., e. and f. a large preference for choosing the target object within the focus area was found for both the focus and the no-focus conditions. In these conditions the right-/left-most target object was always located within the focus area. In the subgroups e. and f., where the target block within the focus area was always identical to the focus-establishing block, the same block was identified twice. Due to the effect of the focus area including the absolute right/left target object within the whole domain, the subgroups b., d., e. and f. were excluded from further analysis.

### 6.5.6 Group 6

Finally, in group 6, a significant result was again found. In the focus condition subjects had a greater preference for choosing the target object out of the focus area than they did in the no-focus condition (t=2.438, p<0.015). This preference is due to the trials in subgroup a. (t=2.336, p<0.0199), where a significant difference was also found between the no-focus condition (12.2% within the focus area) and the focus condition (8% within the focus area).

In subgroup b. no significant difference was found (t=0.706, p<0.4807). Again this is due to the fact that the two blocks involved in the non-focus area were adjacent. Thus subgroup b. was also excluded from further analysis.

### 6.5.7 Summary

In summary, a descending preference for choosing the referent within the focus area in the focus condition was found for an ascending redundancy of the referring expressions (see Table 8). In groups 1 and 2, clear shifts to choosing the object within the focus area occurred in the focus condition. This shift was more prominent in group 1 than in group 2 (resp. 25.8% and 18.2%). In subgroup 4a a shift from choosing the target object outside of the focus area to choosing it within the focus area occurred. This shift was again smaller than in group 2 (12%). Hence, from groups 1 through 2 to 4a a gradually declining positive influence of the focus area occurred.

In subgroups 5a+c, a significant shift (2.7%) was found towards more preference for the target object out of the focus area in the focus condition. Finally, in subgroup 6a, an even larger shift (4.2%) to choose the target object out of the focus area in the focus condition occurred.

Table 8 Summary of the relevant results. Per group, subgroup or subgroups are the percentual differences between the choice for the target object in the focus area in the no-focus and the focus conditions given.

| **group** | **difference -/+ focus (%)** |
|---|---|
| 1 | -25.8 |
| 2 | -18.2 |
| 4a | -12.0 |
| 5a+c | 2.7 |
| 6a | 4.2 |

## 6.6 Discussion

The response pattern that was found in the experiment matched to a large extent the hypotheses that were formulated before. However, some remarks can be made with respect to the data analysis in general, the particular responses in certain groups and subgroups, and the implications of the results for the principle of minimal cooperative total effort and spatial focus.

### 6.6.1 Response times

The data that were recorded during the experiment were the coordinates of the blocks that were touched by the subjects as well as the response times of the subjects. The assumption with respect to the response times was that it would take less time to select a block within the focus area than a block within the non-focus area. However, the response times were not used for analysis, since they turned out to be not very reliable. There was too much variation in the time needed for the physical act of pointing to be reliable for measuring differences in the time needed to decide which block to choose. Moreover, the distances from the location of the hand of the subject to different blocks on the screen were different. Finally, in all probability the different reading times of the various referring expressions due to differences in length influenced response times.

### 6.6.2 The effect of right/left

In all groups where expressions containing 'right/left' or 'right-/left-most' were used (in groups 3, 5 and 6), these expressions urged the subjects to choose the right-/left-most block within the whole domain and not within the focus area only. Use of this strategy by the subjects resulted in a very slight difference or no difference at all between responses in the no-focus and focus conditions. In other words, the impact of the focus area was not strong enough to overrule the effect of interpreting these expressions with respect to the whole domain. Apparently not only the grade of specification of the referring expressions, but also the types of features that were used in the expressions influence the choice for the referent. Particularly in group 3, where ‘right/left’ or ‘right-/left-most’ were the only features used in the expression, no effect of the presence of a focus area was found at all.

In group 5 the preference for choosing the right-/left-most block in the whole domain was not as strong as it was in group 3. In this group the referent within the focus area was always located right-/left-most within this area, but the referent in the non-focus area occupied all three positions alternately. In subgroups a. and b., in the no-focus condition, subjects chose for the right-/left-most block most of the time, but in the cases where the referent in the non-focus area was not located right-/left-most they sometimes chose for the referent in the focus area. In the focus condition, the choice for the referent in the focus area almost disappeared, which indicates that the subjects were indeed affected by the redundancy of the expression.

In the subgroups 5b., d., e. and f. in both conditions subjects preferred equally often the referent within the focus area. In all of these subgroups the right-/left-most referent in the whole domain was also located in the focus area, and in both areas the referent was always located right-/left-most. Hence, it was not possible to choose for a referent in the non-focus area that was located more to the right/left within this area than was the case within the focus area. In these subgroups redundancy of the expression was not strong enough to overrule the interpretation of ‘right/left’ or ‘right/left-most’ as meaning right-/left-most within the whole domain. Perhaps surprisingly, even in subgroups e. and f., where the referent in the focus area was identical to the focus-establishing block, the same strong preference for choosing the referent in the focus area occurred in both conditions. The fact that the same block had to be chosen twice, which was not very common during the experiment, did not motivate the subjects to select the other possible referent.

In subgroup 6a. the same type of response was found as in the subgroups 5a./c. In the no-focus condition sometimes the block that was right-/left-most within the focus area was chosen if the possible referent in the non-focus area was not right-/left-most. However, this was done more often in subgroup 6a. than in subgroups 5a./c. (12.2% versus 5.4%). This can be explained by the fact that in subgroups 5a./c. in one-third of the trials the potential target object in the non-focus area was located right-/left-most, whereas this was the case in only one-fourth of the trials in subgroup 6a. Hence, in subgroup 6a. there were more trials where a preference for the focus area could be expected than there were in subgroups 5a./c.

### 6.6.3 The effect of adjacency

In neither subgroup 4b. nor subgroup 6b. was a difference between the two conditions found: all subjects chose the referent in the non-focus area. Apparently the subjects always interpreted the expression 'to the right/left of' as meaning 'adjacent to the right/ left of'. The adjacent blocks were always located in the non-focus area. An effect of the redundancy of the expression could not be measured here, since in the no-focus condition all referents were already chosen in the non-focus area. What can at least be concluded is that the presence of the focus area did not motivate the subjects to choose the referent *within* this focus area.

### 6.6.4 Combinations of grades of specification

In the experiment six groups were formed containing the six possible combinations of three grades of specification (i.e., ambiguity, minimal specification and redundancy) within the focus area and within the whole domain. From group 1 to group 6 the grades of specification of the expressions within the focus area gradually changed from ambiguous to minimally specified to redundant. Within every grade of specification the same gradual changes occurred with respect to the whole domain.

Beforehand it seemed to be reasonable to assume that the more redundant an expression becomes within the focus area, the bigger the chance that a focus shift to the non-focus area will take place. It was not so clear, however, what the effect of the grade of specification within the whole domain would be on this effect. The implicit working hypothesis was adopted that an increase in specification with respect to the whole domain would lead to a bigger chance that a focus transition would occur. This turned out to be the case. As a post-hoc justification for the appropriateness of the working hypothesis the separate groups are analyzed below.

In group 1, expressions were used that were ambiguous within the focus area; thus they were also ambiguous within the whole domain. In this group, there was an overall shift to choosing the target object within the focus area in the focus condition.

For expressions that are minimally specifying within the focus area, there was still a shift to a preference for choosing the object within the focus area in the focus condition. The preference for choosing the referent within the focus area was expected to be bigger for ambiguous expressions within the whole domain (group 2) than that for minimally specified expressions (group 3). The reason was that, in general, it is more probable that subjects choose a target object within a certain area on the basis of a minimally specified expression than on the basis of an ambiguous expression. Unfortunately, the data obtained in group 3 were not suitable for testing this hypothesis.

For expressions that are redundant within the focus area, a smaller shift to choosing the target object within the focus area or even a shift to choosing the target object within the non-focus area occurred in the focus condition. In the cases where the expression was ambiguous within the whole domain (group 4), the effect of redundancy had to be strong to overrule the ambiguity of the expression within the whole domain. Indeed, there was no preference for the non-focus area, but the preference for the focus area was smaller than it was in group 2. In the cases where the expression was minimally specified within the whole domain (group 5) the urge to select this referent was expected to be stronger than in group 4. Indeed, it was found that there was even a shift to a preference for the non-focus area in the focus condition. In group 6, where the expression was redundant within the whole domain the effect was not very easy to predict. Perhaps this redundancy would prevent the subjects from switching to the non-focus area. This did not appear to be the case, since the results showed that the switch to the non-focus area was even larger than in group 5.

### 6.6.5 Implications for spatial focus and minimal effort

The results provide some evidence for the reality of the existence of a spatial focus area in the blocks domain of the empirical study by Cremers and Beun (1995; chapter 3). In this study humans chose for a suitable object in the area surrounding the object that had been referred to previously, when the referring expression was minimally specifying in this area, but ambiguous in the entire domain. The present experiment has shown that, in this type of situation, subjects indeed prefer to choose an object within this, so-called, focus area.

The results with respect to expressions that were ambiguous or redundant within the focus area can not so easily be translated to the empirical blocks domain. Unlike the participants in the empirical study, the subjects in the experiment were *forced* to make a choice between a number of possible target objects. This resulted in a preference for choosing an object within the focus area if the expression was ambiguous and a preference for an object outside of the focus area if the expression was redundant. In the empirical study participants were allowed to ask for clarification when the expression was ambiguous or when it was not clear whether a focus transition was required or not. A possibility to ask for clarification was not included in the experiment in an attempt to control the responses as much as possible.

Not many requests for clarification occurred in the empirical dialogues, anyway. A reason might be that the participants in these dialogues had more information than just the description of the target object. A referring expression formed part of a larger utterance that was embedded in a dialogue about a task that had to be carried out. This utterance contained, for instance, information about an action that had to be carried out with the target object. This type of information established a *functional focus of attention*, i.e., a set of objects with which the current action could suitably be carried out, of which the target object was an element. This type of information was excluded from the present experiment.

The subjects in the present experiment adhered to the principle of minimal cooperative total effort while interpreting the referring expressions. If the referring expression was ambiguous or minimally specifying within the focus area, they did not bother to look for a possible referent outside of the focus area. They limited the area in which to search for the target object to the focus area, hereby minimizing the expended effort. Only if the referring expression provided more information than was strictly necessary to identify an object within the focus area, they switched to the non-focus area, which evidently takes more effort than staying in the focus area.

## 6.7 Conclusions and future research

In this chapter the results of an experimental study have been reported, that was carried out to test the hypothesis that the presence of a spatial focus area influences the choice of a referent of a certain referring expression. Three different grades of specification of the referring expression, i.e. ambiguous, minimally specified and redundant, were used to investigate in which way they led the subject to choose the referent either within or outside of the focus area. It was hypothesized that the more redundant the referring expression was within the focus area, the more the subjects tended towards choosing the referent outside of the focus area.

In general the hypotheses about the spatial focus of attention and the three different grades of specification are confirmed. A significant difference was found between choices of referents in the condition where no focus area was established and in the condition where a focus area was present. The differences were not the same for the three grades of specification of the referring expressions. As the redundancy of the expression increased, the subjects tended to shift to the focus area less in the focus condition than they did in the no-focus condition. In the trials that contained the most redundant expressions, they tended to shift to the non-focus area.

However, not only the grade of specification of the referring expressions, but also their actual content appeared to influence the choice of a target object. In two particular types of referring expression the lexical contents were interpreted irrespective of the focus area. In response to expressions of the type 'the right/left(-most) one' or 'the right/ left(-most) grey one' the subjects tended to choose the absolute 'right/left' object within the whole domain. Also, in response to expressions of the type 'the grey one to the right/ left of the black one' subjects tended to choose blocks that were located '*adjacent* to the right/left of' the relatum.

Although the results of the experiment seem to be convincing, it is too early to generalize. It has only been shown that in one particular type of domain, of a particular size, containing a particular number and particular types of objects that are placed in a particular arrangement, the hypotheses are confirmed. Further, the referring expressions that were used are of a particular type. In order to be able to generalize the results, other types of domains, objects and referring expressions should be studied.

Nevertheless, the results provide some evidence for the reality of the existence of a spatial focus area in the blocks domain of the empirical study by Cremers and Beun (1995; chapter 3). Furthermore, the subjects clearly adhered to the principle of minimal cooperative total effort while interpreting the referring expressions. If possible, they stayed in the focus area to search for the target object, hereby spending the minimal amount of effort. Only in the case of redundant expressions they assumed that a signal was given to switch to the non-focus area.

The latter finding indicates that overspecification is not necessarily interpreted by the addressee as a help to identify the target object, or to contrast the object with the last one focused on, as Levelt (1989) stated. Overspecification can also carry an additional meaning, for instance, to direct the addressee to look for the target object outside of the current focus area.

# 7 An application domain: object reference in the electron microscope

## 7.1 Introduction

In the previous chapters an attempt was made to analyze the use of referring expressions and gestures in a visually shared domain of conversation. The domain that was taken as a platform for that study is a simple blocks domain. Although the domain was chosen in such a way that it was expected that very basic referring mechanisms would apply, it is not clear to what extent the findings can be generalized to other types of domains. An alternative domain should contain other types of objects, and the relations between the objects and the interaction with the domain (as well as the task) should be of a different type.

The DenK project, which is the context in which the present study was carried out, aims at developing a generic user interface design that should be applicable to any kind of domain. In order to test the current version of the system as well as for demonstration purposes, a specific application domain was chosen. As a result of these considerations an implementation of an electron microscope (EM) was selected as an application domain.

The reasons for choosing the electron microscope (EM) are based on criteria relating to the nature and the goal of the application, the possibilities of modelling the domain, the potential users and the possibilities of interaction with the system (Beun, 1994a). In particular, the EM was considered to be appropriate for external utilization, especially as a training simulator. There is also sufficient knowledge available of the domain, it is rich enough for non-trivial interaction, and it can easily be extended. Further, the intended users of the EM application (e.g., laboratory assistant trainees) have no or little previous experience with the apparatus, and thus form a relatively homogeneous group with respect to domain knowledge. Finally, visualization, direct manipulation as well as natural language play important roles in the interaction with the user.

The EM domain differs from the blocks domain with respect to the types of objects involved, the relations between the objects, and the type of interaction. Therefore, from a theoretical point of view it is interesting to investigate whether the findings in the blocks domain would also hold in the EM domain.

In the following, after having introduced the concept of an application domain in section 7.2, the blocks domain (section 7.3) and the electron microscope domain (section 7.4) are described. The domains are compared with respect to their objects and relations and their respective communicative situations. Furthermore, object reference of the participants as it was found in the blocks domain is described briefly. Based on the findings in the blocks domain, expectations with respect to object reference in the EM domain are formulated.

## 7.2 Application domain and communicative situation

In order to give an adequate description of a certain application domain, it is necessary to distinguish between the domain and the communicative situation in which it plays an essential part (Beun, 1995). This division is also part of the philosophy behind the DenK system.

The domain contains objects with their features, behaviors, relations and interdependencies. In the DenK system, objects are graphically represented on the computer screen. In order to make the domain resemble the 'real world' as closely as possible, the objects are represented three-dimensionally and they may exhibit autonomous behavior.

The communicative situation concerns the kind of interaction that is possible between the user and the domain. The user is a person who possesses certain knowledge and certain intentions, to whom certain modalities of communication are made available. In the DenK system the user can communicate with the cooperative assistant internal to the system by means of natural language and mouse-mediated pointing gestures. For instance, natural language can be used to ask questions about the domain or to give commands to the cooperative assistant. Further, the user can directly manipulate the objects in the domain by means of the mouse. Finally, the user can observe the domain, for instance, in order to inspect the effects of direct manipulation.

## 7.3 The blocks domain

### 7.3.1 Objects and relations

The objects that formed part of the blocks domain were relatively simple objects; namely, Lego® blocks. Most types of blocks occurred more than once in the domain. The blocks did not have names; they could only be distinguished by four different features (color, size, shape and location).

The relations between the blocks were mainly spatial, in three dimensions. Blocks were standing beside each other, on top of each other, etc. The block building was deliberately designed to be as abstract as possible, such that groups of blocks did not form obvious figurative wholes. Consequently, no part-whole relations were expected to occur.

### 7.3.2 The communicative situation

Two studies that represented different communicative situations were carried out in the blocks domain. Both communicative situations included two participants, one of whom gave instructions pertaining to the reconstruction of an existing block building and the other of whom executed these instructions. The instructor had an example of the intended block building available, whereas the executor possessed a collection of blocks.

The communication between the participants was in one study spoken natural language and in the other study natural language that was typed in via a keyboard and read from a computer monitor. The main type of utterance that was used by the instructor was the command. The executor mainly asked for clarification when the command was not clear enough. In both studies the participants were allowed to use gestures that could be combined with natural language, for instance, to refer to objects in the domain.

With respect to the interaction with the blocks domain, in both studies only the executor was allowed to manipulate the blocks. Both participants had visual access to the domain, so they could both observe the gestures and manipulations that occurred in it.

### 7.3.3 Object reference

The process of object reference has already been described in general as consisting of two stages (Levelt, 1982; Cremers and Beun, 1995; chapter 3). In the first stage, the object that is going to be referred to has to be selected in the domain. In the second stage, it has to be decided which referring expression is going to be used to refer to the chosen object.

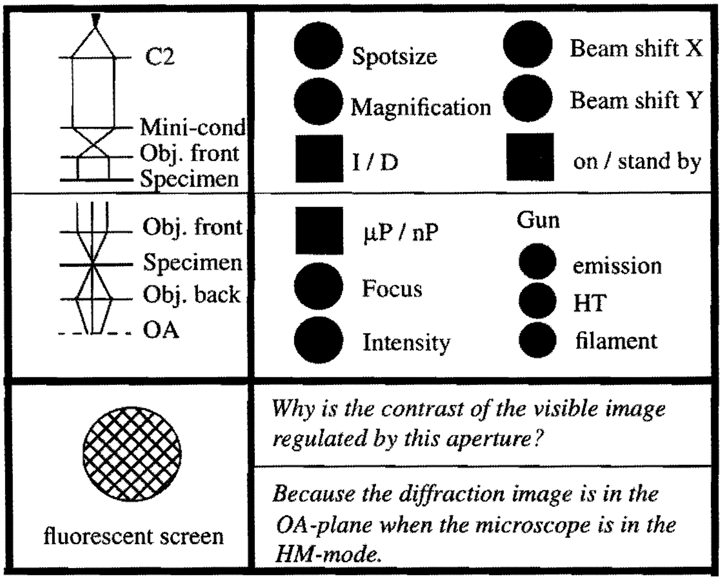
Because the task in the blocks domain required making some specific changes in an existing block building, the instructor knew beforehand which of the blocks were involved. However, the order in which the instructor referred to these blocks was not predetermined. The choice which block to refer to next was decided partly based on the task that had to be carried out and partly on the spatial configuration of the blocks. It turned out that subjects tended to choose as the next object a block located close to the one referred to previously, in other words: within the current spatial focus area. Within the group of nearby blocks they chose the block that was involved in the task, i.e., that had to be manipulated in some way and was accordingly in the functional focus of attention. Hence, the choice criteria were spatial proximity and functional relevance.

Since the blocks did not have unique names, the features of the blocks were used to refer to them. There were a few exceptions: for some blocks that had deviant, non-rectangular shapes sometimes type expressions were invented (e.g., 'the slide', 'the cap'). In general, participants had a preference for using absolute features, in this case mainly color features. The relative features that were used mainly reflected the spatial relations between blocks, for instance, reference to a block by indicating its location with respect to another block ('the red block to the right of the large blue one'). Although care was taken to ensure that the block building was abstract, it could not be avoided that sometimes subjects attached meanings to collections of blocks. For instance, someone called three blocks on top of each other 'the little tower'. For the blocks within this 'tower' a part-whole relationship applied.

There was a difference between the two studies with respect to the length of the expressions that were used. Participants used fewer words in the referring expressions in the keyboard dialogues than they did in the spoken dialogues. For instance, they omitted determiners and used abbreviations for content words.

When the speakers referred to an object close to the previous one they often reduced the information in the referring expression, to indicate that only objects within the current focus area needed to be considered. Speakers also reduced the amount of information in the referring expression if an object was considered to be in the functional focus of attention.

By preferring the use of absolute features and by taking into account the current focus of attention, participants tried to cooperatively minimize the effort expended in referring to an object and identifying it.



*Why is the contrast of the visible image regulated by this aperture?*

*Because the diffraction image is in the OA-plane when the microscope is in the HM-mode.*

Figure 16 A possible layout of the screen representing the electron microscope domain in DenK: EM-scheme (upper left), EM-controls (upper right), EM-image (bottom left), EM-dialogue (bottom right).

## 7.4 The electron microscope domain

### 7.4.1 Objects

A possible layout of the screen on which the EM domain in the DenK system is represented is depicted in Figure 16 (copied from Beun, 1994b). The EM domain consists of a schematic representation of the EM (EM-scheme), a representation of the various controls to adjust the EM (EM-controls), and a representation of the image that is produced by the EM (EM-image). A fourth section of the screen is reserved for the dialogue that is carried out between the user and the cooperative assistant (EM-dialogue). Although the final lay-out of the screen and the objects included in it has not been decided upon yet, for explanatory purposes the components and their positions will be referred to as if they were final.

Most of the objects in the EM domain have unique names, but many of the types of objects occur more than once (e.g., various lenses, various buttons). The electron beam is a special case here. There is only one beam, but it may have different characteristics in various stages of the magnification process. Hence, different parts of the electron beam may be considered to be separate objects.

### 7.4.1.1 A schematic representation of the electron microscope

The schematic structure of the electron microscope is represented in the upper left section of the screen as being approximately the same as that of the traditional light microscope (Beun, 1994b; Beun, 1995; Van Leeuwen, 1993). This structure is transparent, allowing the user to observe the internal state of the EM, in particular the diffraction of the electron beam and the excitation of the lenses. The structure of the EM has also been laid down in a logical model (Ahn, 1995). Two essential differences with the light microscope are that the light source in the EM is not a lamp but an electron gun that projects electrons onto a fluorescent screen, and that instead of optical lenses there are variable electromagnetic lenses.

The basic components of the EM are the electron gun (light source), variable positive electromagnetic lenses (to refract the electron beam), apertures (to cut off specific parts of the electron beam), the specimen and a fluorescent screen (on which the image is projected).

The EM can be in two different *states*: (a) the imaging state, in which an enlarged image of the specimen is projected onto the fluorescent screen, and (b) the diffraction state, in which a diffraction pattern of the crystal structure of the specimen is projected onto the screen.

Furthermore, the EM consists of two different *optical systems*: (a) the illumination system, which illuminates a certain area on the specimen, and (b) the projector system, which projects the plane of interest (see below) on the fluorescent screen.

The *illumination system* is the part of the EM above the specimen. It contains three condenser lenses: the C2 lens, the mini-condenser lens and the objective pre-field lens. The illumination of the specimen can be varied by changing the excitation of the lenses. In Low Magnification (LM) mode the pre-field lens is weak, whereas in High Magnification (HM) mode the pre-field lens is strong. In the latter case the mini-condenser, which regulates the size of the illuminated area, can be either on or off. The on-state is called *microprobe* (μP), the off-state is called *nanoprobe* (nP).

The *projector system* is the part of the EM below the specimen. It contains the objective post-field lens that produces the first intermediate image. The plane where the image is formed is determined by the strength of this lens.

The EM contains three relevant *planes*: (a) the specimen, (b) the objective aperture (OA) and (c) the selected area (SA). An image of the specimen is formed on the plane of the first image. In LM this is the specimen and in HM it is the SA plane. A diffraction pattern of the specimen is formed at the crossover of the electron beam. In LM this is the SA-plane, in HM this is the OA-plane.

The EM contains two *apertures*: (a) the objective aperture (at the OA-plane) and (b) the selected area aperture (at the SA-plane). The apertures select specific information of the image. In image mode this is the part of the image that is projected. In diffraction mode this is the part of the specimen from which the diffraction pattern originates.

The objects that are presented in EM-scheme of Figure 16 can be divided into two groups: the 'hardware' and the 'software'. The hardware of the EM are the parts that are either static or that can be adjusted by means of the controls (i.e., all parts except the electron beams and the specimen). The software of the EM consists of any part that is affected as a result of adjusting the hardware of the EM by operating the controls, or that can be removed from the EM (i.e., the electron beams, the image of the specimen or the specimen itself).

#### 7.4.1.2 The other parts of the electron microscope domain

Beside the schematic representation of the EM in EM-scheme, the EM domain contains two more sections: EM-controls and EM-image (EM-dialogue will be described in section 7.4.3).

EM-controls is located in the upper right section of the screen and contains a representation of the relevant parts of the control panel of the EM. Visually the controls have different shapes (buttons and dials), are placed at different locations, have different labels and have different numbers of states. The non-visual differences of the controls are that they have to be operated in different ways (e.g., push, turn up/down) and that they serve different functions.

EM-image is shown in the lower left section, and contains a representation of the image of the specimen on the fluorescent screen. However, the visual information in the image is not represented in the system. This is due to the fact that the application is only designed for operating the microscope, and not as an expert system about microscopic images.

### 7.4.2 Relations

In EM-scheme there are at least four different types of relations between the objects. In the first place there is a *spatial* relation between all objects that is predominantly linear. It corresponds to the direction in which the electron gun fires the electrons. For example, the objective post-field lens linearly succeeds the objective pre-field lens. There is also a *cause-effect* relation between the hardware and the software components. For example, if the excitation of a lens is changed, then the shape of the electron beam changes. Note that changes in the hardware cannot be induced directly, but that they are always effectuated by adjusting the controls. A third type of relationship between the components is the *part-whole* relationship. The EM consists of several components, for instance, the components above the specimen (the illumination system) and below it (the projection system). The objects that are contained within these systems are their parts. Finally, some *dependency* relations can be distinguished. These are present between two objects if the result of adjusting one object is dependent on the state of the other object. For instance, if the electron gun is turned off, no image will appear on the fluorescent screen.

In EM-control the relations between specific objects within the control panel and with objects outside the control panel are of four different types. In the first place, the objects in the control panel have fixed spatial relationships with each other in two dimensions (e.g., left, below). Secondly, some objects have part-whole relationships with a higher level concept in the sense that they form a group of controls with related functionality. For example, there is a group of controls that applies to only one component of the EM, namely, the gun. Thirdly, all control objects have cause-effect relationships with the hardware parts of EM-scheme that change when a control is adjusted. Indirectly the controls can also adjust the software parts. Finally, some controls have interdependency relations. For example, the controls can only be operated if the 'on/standby' switch is in the 'on' position.

### 7.4.3 The communicative situation

The user can communicate with the cooperative assistant by typing in natural language in EM-dialogue, which is located in the lower right section of the screen. Natural language communication can only pertain to the operation of the EM in general. In other words, only what is represented in either EM-scheme or in EM-controls can be a subject of the dialogue. The user is expected to use two predominant forms of natural language interaction, namely questions and commands. Questions are used to obtain information about objects in the domain and about the operation of the EM. Commands are used to instruct the system to carry out certain operations in the EM.

Furthermore, the user can point at objects or manipulate objects directly by means of the mouse. However, these actions cannot be carried out in all sections. In EM-scheme only pointing actions can be used. In EM-control both direct manipulation and pointing actions can be carried out. All of the objects in the control panel can be adjusted in some way. In EM-image no interaction can be carried out at all.

A special remark must be made here about the effects of directly manipulating the control buttons. If an adjustment of the controls is carried out, effects may occur in EM-control, in EM-scheme and in EM-image. The effect in EM-control is that a button is pushed, or a dial is turned up or down. The indirect effect of this manipulation can be observed in EM-scheme and, if the resulting image is affected, in EM-image as well.

A final way of interaction is that the user can observe EM-scheme, EM-controls and EM-image. It is possible for the user to see the immediate effect of, for instance, adjusting a control or giving a command to the assistant. It is even possible to see more than in a real EM, since the EM is transparent here, making the electron beam and the various lenses visible.

### 7.4.4 Object choice and reference

#### 7.4.4.1 Choice of an object

In the blocks domain the order in which the blocks were referred to was partly based on the task and partly on the spatial configuration of the blocks.

The order in which the objects are referred to in the EM domain depends on the intentions of the user. If the user just wants to explore the workings of the EM, it is not easy to predict how the user will proceed in gathering information. In this situation the choice of the next object to refer to may be spatially determined. The user may start by exploring the apparatus by asking for the names and/or functions of the components in a systematic way. If the user actually wants to operate the EM, the choice of the object to refer to next is expected to be mainly functionally determined. In this situation it may be possible to predict in which order the different controls will be manipulated in order to achieve a certain result.

### 7.4.4.2 Reference to an object

**Features**

In the blocks domain mainly references to physical features of objects were used, with a preference for absolute features. On the one hand, the objects in the EM domain are more complicated and more distinct than those in the blocks domain. Also, fewer identical objects occur in this domain. Hence, the number of features that can be used to refer to them is not as limited as in the blocks domain. On the other hand, most of the objects in the EM domain have unique names that can be considered absolute features and that can be used for reference.

Considering the preference for using absolute features in the blocks domain, this would predict that in the EM domain absolute unique names will be used, e.g., 'the mini-condenser'. This may seem the easiest way of referring, but in many cases the names are rather long and/or complex. In order to reduce effort users will most probably try to type in as few words as possible. If the names are long, the users are expected to omit determiners and use abbreviations (as was also the case in the keyboard dialogues concerning the blocks domain). They may also type in features instead of names, but only if this takes less effort.

Contrary to the objects in the blocks domain, the objects in the EM domain have not only spatial relations with respect to each other, but also part-whole, cause-effect and dependency relations. This suggests that if relative features that reflect relations between objects are used, they will not only concern spatial relations. Again, in a situation where users are exploring the domain, more use will probably be made of spatial relations. In this situation expressions such as 'the button to its right', where 'its' refers to the button referred to previously, are expected.

All other relations between objects that occur in the EM domain can in principle be used to refer to an individual object.

The part-whole relation can be used to refer to an object that forms part of a larger component, for instance, in 'the lens in the projection system'.

The cause-effect relation can be used to indicate an object that is either a cause or an effect of a certain change in the domain. Two examples are 'the lens that changes this part of the beam if it is adjusted' and 'the part of the beam that changes if the objective post-field lens is adjusted'.

The dependency relation can be used to indicate an object whose features depend on the state of another object, for instance, in the expression 'the image in diffraction mode'. However, the last example is different from the previous ones. This expression is referring to a hypothetical state that is not present at the moment of the utterance, but that may have been the case in the past or may become the case in the future in the interaction. In the classification of the types of referring expressions that occurred in the blocks domain presented in chapter 2, this type of expression did not occur. In the framework of the construction task there was probably no use for speaking about hypothetical states. Expressions that referred to the history of an object did occur, though (e.g., 'the block that just fell off').

**Reduced reference**

In principle, all objects in the EM domain can be referred to uniquely, by means of a unique name or a unique combination of features (including the location of the object). Since the names of many parts are relatively complicated and/or long and the natural language communication has to be carried out by means of typing, it is expected that users will prefer to use a shorter and/or simpler expression, for instance 'the lens'. However, since many of the objects also have features in common, including the various lenses, parts of the electron beam, planes, apertures, buttons and dials, such an expression would be ambiguous if the entire domain were to be considered. This is the reason why users are expected to take into account the current spatial and functional focusses of attention, as well as other types of relations that occur between objects in the EM domain.

Spatial focus reduction is expected to occur in situations where the user's choice of the next object to refer to is spatially determined, for instance in the exploration phase. An example is a situation in which a user is exploring EM-scheme. When the user has just asked about the functions of the C2 lens and the mini-condensor lens, and subsequently asks for the function of 'the next lens', its is very likely that the objective prefield lens is meant.

Functional focus reduction can be used in the EM domain by means of naming the action that has to be carried out with a certain object. This seems to especially apply in EM-control. In this section the controls can not only be referred to by mentioning their names, but also by mentioning the operation that has to be carried out. An example of such an expression is: 'Put the button in diffraction mode'. Although the control panel contains three buttons, only one of them can be put in diffraction mode, so the expression is unambiguous.

In the same way the part-whole relation can be used for reducing the expression. For example, if the user indicates that the projector system is under consideration at the moment and the referring expression 'the lens' is used, then only the lens in this system, i.e., the objective post-field lens, can be the referent.

The cause-effect relation can be used for reduced reference in yet another way. For instance, if the referring expression 'the electron beam' is used in an expression such as 'if the magnification of the mini-condenser lens increases, then what happens to the electron beam?', the cause-effect relation will predict that the part of the electron beam that is meant here is the part leaving the mini-condenser lens. Neither the spatial nor the part-whole relation would help in solving the reference here. Spatially the incoming and the outgoing beams are equally close to the mini-condenser, and they both form part of the illumination system of the EM.

The dependency relation may also result in reduced reference that can not be solved by any of the other types of relations. This is the type of reference where the user assumes that the system knows that a dependency relation between two objects exists and uses this knowledge as a disambiguating device. An example of such a reference is included in the following question: 'How is the image formed on the plane in HM mode?'. Provided that no previous reference to a plane has been made, the expression 'the plane' is not ambiguous here, since in HM mode an image is only formed at the SA plane.

**Pointing and direct manipulation**

In the above it is assumed that users who interact with the EM domain will only use natural language expressions. But of course there is also the possibility of pointing at objects or manipulating them directly.

For instance, it takes less effort to point at objects by using the mouse than to type in a long referring expression. Also, in many cases it will take less effort for a user to directly manipulate a control than to order the cooperative assistant to do this.

However, the preference for pointing at an object instead of using natural language also depends on the characteristics of the object involved as well as the number of objects the user intends to refer to.

It is to be expected that the user will point especially at the objects that can easily be distinguished from their surroundings, are not part of a larger object, and do not have other objects included in them, for instance, the controls. The controls are the only objects that can be manipulated directly.

Objects that form part of a larger object or set of objects (for instance, any part of the EM) or include other objects (for instance, the illumination system) are expected to be referred to more often by means of natural language, possibly combined with a pointing gesture. The reason is that in these cases there is always an ambiguity concerning which object is meant by the pointing gesture alone.

Finally, in cases where the user wants to refer to a set of objects that have certain features in common, it is probably easier to use quantification in natural language than to point at all objects one by one (e.g., 'all controls').

## 7.5 Conclusions

In conclusion, users of the EM application domain will make use of the characteristics of this domain and its interaction possibilities to minimize the effort needed to refer to objects. In principle, they have a wide range of possibilities available to do this. They can use pointing gestures and direct manipulation if the object is suitable for that. If they use natural language, they will strive to reduce the referring expressions used as much as possible. Since the characteristics of the EM domain differ a lot from those of the blocks domain, the way in which minimization of effort is accomplished will be different too.

The most conspicuous difference between the two domains is probably that objects in the EM domain have unique names, which is not the case in the blocks domain. This means that it is always possible to refer to an object by using its name, or to come to an agreement with the system to use an abbreviation of the name if it is too long. It is less likely that users of the EM domain will use combinations of features of the objects to refer to them, which was very common in the blocks domain.

The order in which the objects are expected to be referred to in the EM domain is functionally rather than spatially determined. This functional perspective has also repercussions for the use of the spatial and the functional focusses of attention. Spatial focus is expected to be less important in the EM domain than it is in the blocks domain. It is to be expected that functional focus will be made use of more often, in particular when the user is instructing the system to operate the controls.

The objects in the blocks domain only have spatial relationships with each other, whereas the objects in the EM domain also have other types of relationships (part-whole, cause-effect and dependency). These relationships are probably more prominent for the user than the spatial ones, since they are directly related to the design and the functionality of the EM. They are expected to be more prominent than spatial relations in object references.

Pointing was used a lot in the blocks domain, especially in the keyboard dialogues, where it took a lot of effort to type in referring expressions. Although the user of the EM application domain also has to type in natural language expressions, pointing is expected to take place less often here than in the blocks domain. In the EM domain it is not clear whether pointing will be effective, since, generally, objects are harder to distinguish from each other than in the blocks domain.

In order to test the predictions that have been formulated in this chapter, 'Wizard of Oz' experiments should be carried out to study the actual referring behavior of users that are interacting with the EM domain. By analyzing the dialogues that arise from these interactions, it can be decided whether, for instance, it is essential for the system to incorporate notions of the spatial and functional focusses of attention in this type of domain.

# 8 A comparison of the DenK system with other multimedia systems

## 8.1 Introduction

In recent years many attempts have been made to develop human-computer interfaces that allow interaction by means of a combination of graphical and natural language (multi-modal) communication. The question can be asked why it was considered worthwhile to develop yet another one of these types of systems: the DenK system.

It is claimed that an important distinguishing characteristic of DenK is that its communicative situation reflects the 'natural' situation in which two dialogue participants communicate when they cooperatively carry out a certain task in a shared domain. Key elements in this type of communication are that *symbolic interaction* (natural language) and *physical interaction* (direct manipulation and pointing gestures) can be carried out. Furthermore, the domain is separate from the dialogue participants (it is actually the 'real world'). It exists in its own right, and is not, for instance, just represented in the minds of the participants.

In this chapter, first a general overview is given of seven multimodal systems. The systems are then compared with respect to both their abilities to generate and interpret referring expressions and gestures and the extent to which they incorporate a notion of the spatial and functional focusses of attention. Finally, the systems are compared to the DenK system, in particular with respect to their communicative situations and the way in which reference and focusses are treated.

## 8.2 Aspects to compare

### 8.2.1 General

The systems that are presented here differ in many respects, of which some are discussed below. They are partly based on a checklist of aspects of dialogue (Mason and Edwards, 1988) and a taxonomy of dialogue systems (Van Deemter, Beun and De Vet, 1995).

The systems that are chosen to be discussed here all deal with *human-machine dialogues*, i.e., the type of dialogue where a user is communicating with a system. Further, in all systems the purpose of the dialogue is in some way *task-oriented.* The systems allow the user to command the system either to carry out some *action*, issue an *information query* to get some information from the system, or both. Some of the systems are in a stage of development where only a rudimentary dialogue component exists. These systems just provide information, thereby assuming that an information query has been carried out by the user.

The systems all have specific *application domains*, each having different characteristics. The *modes of interaction* between the user and the system (including the application domain) are of many different kinds:

1. Natural language (NL) (spoken and/or written).
2. Gestures (in particular mouse-mediated pointing by the user, and cursor-mediated pointing or highlighting by the system).
3. Command languages (CL).
4. Direct manipulation (DM) of graphical objects (mouse-mediated).
5. Menu selection (MS).
6. Form-filling (FF).
7. Combinations.

Not all interaction modes apply to both the system and the user. Natural language and gestures can be used by the participants. Command languages and direct manipulation can only be applied by the user. In the menu selection and form-filling modes, the system and the user each interact in a different way. The system presents a menu or a form, whereas the user selects an item from the menu or fills out the form, respectively. In particular the latter interaction mode puts the initiative with the system, whereas the former is user-initiated.

Further, the systems use different *models* to represent particular knowledge that is represented in the system. This includes general knowledge about the world, knowledge about the domain of conversation, the user and the ongoing dialogue. The systems use different types of *knowledge representation* methods. Finally, some systems use *external information sources* to supply their own internal knowledge, usually the domain knowledge.

### 8.2.2 Reference and focus of attention

An important way in which the available modalities of communication can be used is for both the user and the system to refer to graphical objects that appear on the screens of the systems. To refer to these objects, the available knowledge bases usually have to be used in order to select the right expression.

Since previous research has shown that the spatial and functional focusses of attention are important concepts in the generation and interpretation of referring expressions, the systems are screened for provisions to deal with these concepts, for instance, knowledge bases that keep track of the objects that are in the focus of attention.

## 8.3 The systems

In the following, seven graphical and natural language interfaces (WIP, COMET, ALFresco, EDWARD, XTRA, AIMI and CUBRICON) are described, mainly in terms of the aspects described above. An overview of all the systems (including DenK) in terms of the most relevant general aspects is provided in Table 8. In Table 10, an overview of the use of deictic reference and the possible incorporation of a notion of focus in the systems is given.

Table 9 Aspects of the systems: the name and the language used, the application domain, the input modes, the output modes, the models of knowledge within the system, the type of knowledge representation and the external information sources that are used.

| system name (language) | application domain | input modes | output modes | knowledge and models | knowledge representation | information sources |
|---|---|---|---|---|---|---|
| WIP (German/English) | physical devices | x | NL (text), graphics, screen/hard copy | geometry, domain, presentation techniques and strategies | RAT | x |
| COMET (English) | military equipment | MS | NL (text), graphics | content and media-specific knowledge | schema-based | expert system |
| ALFresco (Italian) | Italian frescoes | NL (text), pointing, hypermedia buttons | NL (text), film, images, zooming in, hypermedia buttons | domain knowledge, user's interest, topic module | based on KL-ONE, activation network | video disc, hypermedia network |
| EDWARD (Dutch) | file system | NL (text), pointing, DM, MS, CL | NL (text), pointing, reverse video | domain knowledge, context model (salience) | based on KL-ONE | x |
| XTRA (German) | income tax form | NL (text), pointing, MS | NL (text), pointing, FF (tax form) | factual knowledge given by user, dialogue memory, focus structure, user model | SB-ONE | expert system |
| AIMI (English) | mission planning | NL (text), different types of pointing, DM | NL (text), pointing, maps, charts, images, non-speech audio | domain knowledge, context (intentional, attentional) incl. non-ling., presentation | KL-ONE | expert system (MAC-PLAN) |
| CUBRICON (English) | US Air Force | NL (text, speech), pointing | NL (text, speech), pictures, forms, tables, histograms | world/domain knowledge, presentation planning, attentional focus, user model | SNePS | x |
| DenK (English) | electron microscope | NL (text), pointing, DM | CTT (formulas), highlighting, *(graphics)* | private & mutual knowledge, dialogue context, *(domain)* | CTT | x |

Table 10 Provisions for the generation and interpretation of deictic references in the systems.

| system | deictic reference (linguistic and *non-linguistic*) | | focus of attention |
|---|---|---|---|
| | generation | interpretation | |
| WIP | definite expression with spatial location, arrow indicating component, short label with line to component | x | zoomed-in view of a component of the device |
| COMET | definite expression, highlighting | x | shift of the ‘camera position’ to center the target object |
| ALFresco | definite expression | combination of modalities (natural language and pointing) to resolve ambiguities | spatial focus area=current fresco, zooming in to an area |
| EDWARD | referring expression: distinguishing info. (knowl. base), salience (context model), visibility of the referent, pointing (arrow) | combination of modalities, salience of potential target objects | spatial focus area=salience, visible objects |
| XTRA | referring expression, pointing hand (depends on type of object) | combination of modalities (different types of gestures) | spatial focus area=tax form, functional focus: case-frame analysis |
| AIMI | referring expression, highlighting (depends on user’s intention) | no special provisions | spatial focus=whole display |
| CUBRICON | referring expression, blinking/highlighting (visibility of the referent) | combination of modalities | spatial and functional focus (ambiguous pointing) |
| DenK | CTT formulas (mutual knowledge), highlighting | inspect domain | spatial focus area=window |

### 8.3.1 WIP

#### 8.3.1.1 General aspects

WIP (Wahlster et al., 1983) is a system that generates multimodal presentations with instructions for assembling, using, maintaining or repairing certain physical devices. At present the domains of using an espresso machine, assembling a lawn mower and installing a modem have been implemented.

At the present stage of development WIP does not allow the user to interact with the system. Only the expertise level (novice or expert), the target language (German or English), the layout format of the presentation (screen or hard copy) and the output mode (incremental versus complete output) can be specified by the user. A future plan is to include the possibility for the user to interact with the system during the presentation in order to be able to alter the presentation or ask questions about it.

The system generates a simple 3-D wire-frame graphical representation of the physical device and presents it on the screen along with coordinated instructional text.

WIP uses several models to generate the presentation. The first one is an analogical representation of the geometry of the device. The second one is an application knowledge base containing the ontology and abstract plans for a particular domain. Finally, WIP includes a knowledge base containing common-sense knowledge of presentation techniques and strategies, which is used to select the optimal mode of presentation for a certain piece of information. For example, quantification is presented in the form of text, whereas information about visual attributes of objects is preferably presented in graphics.

The knowledge bases are represented in the hybrid knowledge representation system Representation of Actions in Terminological logics (RAT).

#### 8.3.1.2 Reference and focus of attention

The text and graphics that are produced by WIP have to be coordinated so that they form a coherent package of information. In particular the referential relationships in one modality to presentations in another modality have to be established. To do this the graphics generator computes a spatial relation describing the absolute location of an icon in the picture, and this relation is then expressed by the text generator. This expression may be accompanied by an indication in the graphics of which part is meant. For example, in the case of the espresso machine, the text 'open the lid' is supplemented by an arrow in the picture showing in what direction the lid must be opened.

As a result of the demand to create an unambiguous referring expression, various levels of recursion may occur. For instance, different frames of reference may have to be used in one expression in order to denote a unique referent, e.g., in 'in the upper left part of the figure on the right'. If expressions become too long to fit in the picture or if too much recursion in the description is needed, the system offers labelling techniques. A short label is placed in the figure and connected to the corresponding icon, and the same expression is used to refer to this icon in the accompanying text.

The focus of attention of the user is directed, for instance, by providing a more detailed, zoomed-in view of a certain component of the device.

## 8.3.2 COMET

### 8.3.2.1 General aspects

COMET (Coordinated Multimedia Explanation Testbed) (Feiner and McKeown, 1990) provides instructions for equipment maintenance and repair. The system is an interface to an expert system on maintenance and repair of military equipment (for instance, a radio transmitter) for the US Army.

Users (military technicians) can only interact with the system by means of a menu through which the expert system can be invoked and requests for explanations can be made. The system generates both 3-D graphics of the equipment and accompanying textual explanations.

COMET has access to three main knowledge bases when generating a presentation. The external expert system is consulted to determine which problems the equipment is experiencing, which components are suspect and which tests would be most useful in identifying the causes. The domain knowledge base is consulted to determine which type of information should be included in the explanation. The schemes that are produced contain the full content for the explanation in a media-independent form. They can contain six different types of information: location, physical attributes, simple actions, compound actions, conditionals and abstract actions. The media-specific knowledge base is consulted to decide on the media in which particular types of information are to be presented. For instance, location information is preferably represented in graphics, and conditionals in text, while compound actions can be expressed in both text and graphics.

### 8.3.2.2 Reference and focus of attention

The text and graphics generators in COMET interact bidirectionally in order to produce tightly integrated explanations. Bidirectional interaction is, among other things, necessary for producing appropriate referring expressions in the text to refer to components in the graphics. The textual counterpart of reference is the production of a definite noun phrase. The graphical counterparts of reference are the highlighting of the component and a shift of the component to place it central in the picture. Bidirectional interaction makes it possible to adjust the content of the definite noun phrase to changing circumstances in the graphics. For instance, the moment a red dial becomes highlighted, the definite expression 'the red dial' may change to 'the highlighted dial'.

The highlighting of a component and the shifting of it to a central position in the picture are means of attracting the attention of the user. In doing this, less information is needed in the accompanying referring expression.

COMET's media-specific knowledge is used in generating the reference. The location information about the component could also have been expressed in text by means of a prepositional phrase, but the media-specific knowledge calls for expressing it in the graphics part.

### 8.3.3 ALFresco

#### 8.3.3.1 General aspects

ALFresco (Stock, 1991) is an interactive system that supplies information on Italian frescoes. The system is connected to a video-disc unit and provides access to a hypermedia network. The video-disc includes images and film fragments of Fourteenth Century Italian frescoes and relevant monuments. The hypermedia network contains unformalized knowledge, such as critics' opinions about the frescoes.

The user can ask questions by typing in natural language (Italian) about frescoes and related topics. It is also possible to refer to certain designated areas of a displayed image of a fresco by touching them (the system contains a touch screen). Hypermedia buttons make it possible for the user to enter the hypermedia system in order to receive more information about that particular fresco or to start browsing through the system.

The system can show images or film fragments, and it can give answers to questions posed in natural language by the user, by replying with instances such as the title and location of a fresco. It can also give a more complex description of a fresco through natural language. Texts and images are usually accompanied by hypermedia buttons. Also, the dialogue may lead the system to automatically zoom in to details of a displayed fresco or display other frescoes that are in some way related to the one currently being shown.

ALFresco contains three main knowledge sources: domain knowledge, a user's interest model and a topic module. They are all consulted by a pragmatic component to decide how to react in a given dialogue situation.

The domain knowledge is represented in a knowledge base and in the hypermedia network. The information in the knowledge base is expressed in a language that is based on KL-ONE (Brachman and Schmolze, 1985). It defines everything the system can reason about, such as frescoes, monuments, painters and towns that inhabit the frescoes.

The user's interest model is represented in the form of an activation network in which each node represents a particular concept. The connections between the nodes are based on pragmatic closeness of the concepts on the basis of three dimensions of interest: art schools, towns and time periods. Every time a certain concept is referred to by either the user or the system the corresponding node is activated. The generation component has a preference for selecting the most highly activated information.

ALFresco's topic module contains a stack of dialogue turns and a stack of topic units. The former stack contains all the referents that are generated during one turn, either by the system or by the user. The topic units that form part of the latter stack contain all the referents that are produced while discussing one particular fresco.

#### 8.3.3.2 Reference and focus of attention

In ALFresco the user can refer to objects by means of typing in a natural language referring expression and/or pointing by touching the screen.

Natural language can be used to ask questions about frescoes, possibly accompanied by pointing to a certain designated entity that forms part of a fresco being displayed. ALFresco associates the accessible entities with the regions they occupy on the screen. The system is able to solve ambiguity in one of the modalities used with the information provided in the other modality. The deictic context changes with every fresco. Non-deictic references are solved by the system by means of the topic module, by assuming that the user normally refers to entities bound to the current fresco.

Further, the images and texts that are displayed by ALFresco usually contain hypermedia buttons that the user can click on to receive more information on a particular topic or to enter the hypermedia network.

ALFresco assumes that the fresco that is currently shown is the current spatial focus area (the deictic context). Further, it is possible to zoom in to a certain part of the fresco, to delimit the size of the focus area.

## 8.3.4 EDWARD

### 8.3.4.1 General aspects

EDWARD (Bos, Huls and Claassen, 1994; Claassen, 1992; Huls, Bos and Claassen, 1995) is a research prototype of a generic multimodal user interface. The system combines a graphics editor with a Dutch natural language dialogue system.

The current application domain is a file system environment containing a hierarchical tree structure of labelled file icons, a garbage container, a copying machine and an icon picturing the bear Edward, who impersonates the system.

The approach that is taken for the possible modes of interaction with the system is called fully integrated multimodality (Bos, Huls and Claassen, 1994). This means that all modes of interaction are offered at the same time, so the choice of which modality or combination of modalities to use is entirely up to the user. The user can make changes in the tree structure and search for, make copies of, delete or get information about particular files in four ways: (1) by directly manipulating selected objects, using a mouse, (2) by selecting actions from menus, (3) by entering commands in a command language, and (4) by typing a command or question in natural language. Combinations of either commands or natural language with pointing arealso allowed.

The graphical generation component of the system is the file system environment. EDWARD further answers questions in written natural language, possibly accompanied by pointing gestures carried out by the bear Edward. Inverse video is used to indicate icons that have been selected by the user.

EDWARD has two main knowledge sources. The first one is a hierarchical semantic network based on KL-ONE (Brachman and Schmolze, 1985), in which classes and instances of domain entities and relations are represented. The second knowledge source is a context model, which was developed by Alshawi (1987), that is based on the notion of salience. This context model is used to interpret and generate referring expressions.

### 8.3.4.2 Reference and focus of attention

The referring actions that can be interpreted and generated in EDWARD are pointing gestures, and unimodal and multimodal referring expressions (referring expressions accompanied by pointing gestures) (Claassen, 1992).

The KL-ONE knowledge base and Alshawi's (1987) context model are used to generate references. The knowledge base is used for finding information to distinguish the intended referent from other potential referents. The context model is used for determining the salience of referents. Salience depends on linguistic and perceptual factors (including visibility, selected by the user, and indicated by Edward). Depending on the salience value of the referent, the availability of distinguishing information about the referent and the presence of other salient objects, either some kind of referring expression or pointing gesture is generated.

EDWARD's pointing gesture is an arrow that extends from the bear Edward to the target object and then shrinks back again. Subsequently four small arrows are displayed around the object such that they are pointing at it. These arrows are mainly used as a first response to a request in order to give feedback about the selected referent.

EDWARD interprets unimodal linguistic reference in the following way. If the user uses a definite expression referring to a unique object of a certain type, then the most salient object of this type is selected. All linguistic expressions describing spatial relations are interpreted deictically. To determine the referent of a spatial expression, the domain is scanned for a referent. Multimodal references are in principle interpreted in the same way as unimodal referring expressions. If the pointing gesture is ambiguous it is interpreted by considering the accompanying referring expression (for instance, in the case of 'this book ↗', only books are selected). EDWARD's feedback to unimodal graphical references (mouse-mediated pointing) is to mark the indicated object by means of reverse video to show that it has been selected.

The focus of attention is incorporated in the notion of salience and in the visibility of objects. In generation, salient objects are referred to differently than non-salient and non-visible objects. In interpretation, the objects with the highest salience value are selected as the referent for a definite expression. The visibility of a referent determines whether it is in the current focus area and, whether, for instance, it can be pointed at. Which part of the domain is scanned for finding the referent may depend on the salience of objects. For instance, if a particular directory is very salient, the referring expression 'the file on the left' is interpreted as 'the file on the left in the salient directory'.

## 8.3.5 XTRA

### 8.3.5.1 General aspects

XTRA (eXpert TRAnslator) (Allgayer et al. 1989; Reithinger, 1987; Schmauks and Reithinger, 1988) is a generic cooperative access system to expert systems. At present the system is being applied to the income tax domain. A German income tax form is represented on the screen and the user can be assisted by the system in filling out the form.

To initiate the interaction the user can try to fill out the form by typing in the information that is required. If problems arise questions can be asked to the system by typing in natural language (German), possibly combined by a mouse-mediated pointing gesture to a certain entity or an entry in the form.

XTRA's point of departure for communication is always the tax form that is displayed on the screen. Communication takes place in written natural language and by gestures that are simulated by an icon representing a pointing hand. The system answers queries as to terminology and provides user-accommodated natural-language verbalizations of results and explanations provided by an expert system on income tax.

The system contains a conceptual knowledge base that is represented in SB-ONE (a variant of KL-ONE (Brachman and Schmolze, 1985)). This knowledge base primarily contains factual information that has been provided by the user. The tax expert system contains all domain-specific knowledge. Further, XTRA keeps a record of dialogue contributions in the dialogue memory and a record of the currently focussed objects in the focus structure. Finally, the system maintains a user model with assumptions concerning the user's goals, knowledge and beliefs (both true and false).

#### 8.3.5.2 Reference and focus of attention

XTRA can generate deictic referring expressions accompanied by gestures (Reithinger, 1987). The type of expression or gesture that is used depends on both the type of 'object' that is being referred to (a region of the form, an entry in a value region or a concept of which the region is an instance) and the context in which it occurs (in a sequence of references, as a contrast to another object).

Each object or event that is introduced during the dialogue, as well as the objects that are visible on the screen, is added to the set of referential objects. These referential objects are connected to their occurrences in the dialogue, which is represented in the dialogue memory. Pronouns are used by the system to refer to objects or events that were used in the current context, which is the current or the previous sentence. Definite descriptions are used in the other cases.

The user can ask questions in natural language and may accompany them with pointing gestures. In order to carry out a gesture the user has to choose from a menu that offers four options for pointing: exact pointing (e.g., by means of a pencil), standard pointing (by means of an index finger), vague pointing (by using the entire hand) and encircling larger areas. The resulting pointing action can be ambiguous, since it can refer to a basic region in the form, to the actual content of the region, to a super-ordinated area, or to a concept related to one of the mentioned areas.

The meaning of a pointing expression is interpreted by the system according to the context in which it occurs, especially with respect to the expression it accompanies. Also a case-frame analysis can be carried out that selects the most probable referent on the basis of the required manipulation. For instance, if the accompanying expression indicates that a certain amount should be added to the referent, then the referent can only be contained in a certain region indicated by the pointing act.

The case-frame analysis is a form of functional focus. The spatial focus area is always the whole tax form. The system can direct the focus to a certain region of the form by pointing at it.

### 8.3.6 AIMI

#### 8.3.6.1 General aspects

AIMI (An Intelligent Multimedia Interface) (Burger and Marshall, 1993) is a portable multimedia/multimodal interface to the application domain of mission planning, which assists users in devising transportation schedules and routes. The application domain is an expert system that is called MACPLAN (Military Airlift Command Planner).

The user can ask questions by typing in natural language (English). All textual items in the response, such as table entries and labels in bar charts, are mouse-sensitive. The graphical elements are also partly directly manipulable. For instance, an arrow that links two cities on a map can be relocated to, e.g., link the city of departure with another city of destination.

AIMI's modalities of communication include natural language, maps, highlighting, business charts, still images and non-speech audio. The design of graphical displays is automated. The generation of natural language includes simple labels for data points, chart axes and legends, as well as the presentation of answers to questions of the user in full text.

AIMI's knowledge is represented in the form of KL-ONE (Brachman and Schmolze, 1985) types of predicate subsumption hierarchies that are independent of the modality in which the information is ultimately presented. AIMI also uses two different context models that are based on Grosz and Sidner's (1986) work on discourse structure. The first one is an intentional model, in which intentions of the users are represented. These are taken into account when AIMI presents responses. The second one is an attentional model, in which entities that are in the current focus of attention are stored. This model is used to resolve referring expressions. Both models also incorporate non-linguistic actions. Before presenting the information, the system has decided which modality is most appropriate for presenting the information in. In choosing the type and modality of information the system takes into account the goal of the user and the context in which a request is being made.

#### 8.3.6.2 Reference and focus of attention

In AIMI, the user can click on textual items that occur in a graphical display and ask questions about them in natural language, e.g., '[User clicks on the table entry F-4C] What is its speed?'. Many graphical displays contain 'Text' buttons that can be clicked on, and which cause the system to display textual information about the contents of the display. The user is also allowed to refer to elements of the interface itself, such as charts and maps. If the user clicks on an entity or refers to it by means of natural language, the entity is introduced in the context. The system resolves later references to it, using the attentional context mechanisms.

In AIMI, the focus of attention is only used to resolve anaphoric expressions. The whole graphical display is considered to be the spatial focus area. There is no way for the system to direct the user's attention to a smaller area. The user can direct the system's attention by pointing (clicking on items).

### 8.3.7 CUBRICON

#### 8.3.7.1 General aspects

CUBRICON (CUBR Intelligent CONversationalist) (Neal and Shapiro, 1991) is a knowledge-based multimedia interface system. Its application domain is the US Air Force Command and Control.

The user can communicate by means of spoken and written natural language (English), and may choose to accompany this with a mouse-mediated pointing gesture.

The system can produce color graphics/pictorial displays (such as geographical maps), tables, histograms, spoken and written natural language and forms to be filled out. The general premise is that a graphical presentation is always desirable. Tables and histograms are used if the information to be presented lends itself to either of these types of presentation media. A form is used if the task requires it. Spoken natural language is used to give explanations on graphics, warnings, information on the system's activity and short non-technical expressions (e.g., yes/no answers).

The system has knowledge bases of output planning strategies, of world knowledge and of domain knowledge. The knowledge is represented in the SNePS semantic network processing system (Shapiro, 1979).

CUBRICON also builds up a discourse model consisting of an attentional discourse focus list that is based on Grosz and Sidner's (1986) attentional focus. This model consists of a main focus list and a display model. The main focus list contains the natural language referring expressions that have been explicitly expressed, pointing gestures, highlighting and blinking. The display model contains all visible objects on the screen.

Finally, CUBRICON maintains a user model that contains information about the types of entities that are considered important by the user and about the task in which the user is engaged. It is used to determine what information is relevant in answering questions or responding to commands, and to determine in which form the information should be presented.

#### 8.3.7.2 Reference and focus of attention

CUBRICON uses its discourse model to interpret and to generate written and spoken referring expressions and pointing gestures. For interpretation, the referent of a pronominal expression is searched for in the main focus list. If a definite expression is used, first a referent is searched for in the main focus list. If no suitable referent exists there, the display model is inspected. If more than one suitable referent exists there, all these referents are selected. A mechanism should be added that chooses the referent that is most relevant to the user's task, but this has not as yet been implemented.

To generate a reference a deictic dual-media expression (natural language and blinking/highlighting) is used if the referent is part of the display model (i.e., if it is visible). If the referent is the most salient entity in the main focus list, a pronoun is used.

The user can use coordinated natural language (either spoken or typed) and pointing. There are four types of objects that can be pointed at: geometric points, icons, table entries and windows. Problems may arise if the pointing gesture is directed at an overlapping area of two or more objects or if the intended area is missed altogether. For instance, if the area pointed to contains no instance of the object that is expressed in the accompanying natural language expression, a spatial search is carried out to find the referent. If more than one potential referent is present in the indicated area, a referent is searched for that either has the property that is expressed in the accompanying expression or with which the task can be most suitably carried out (the task is represented in CUBRICON's user model).

In CUBRICON, the attentional focus of attention (Grosz and Sidner, 1986) is used to resolve anaphoric expressions. The spatial focus of attention is applied when the system looks for the referent in the vicinity of the location at which the user has pointed. The functional focus is applied if the task to be carried out is taken into account. The system may attract the attention of the user by highlighting or blinking a certain object, if it is visible.

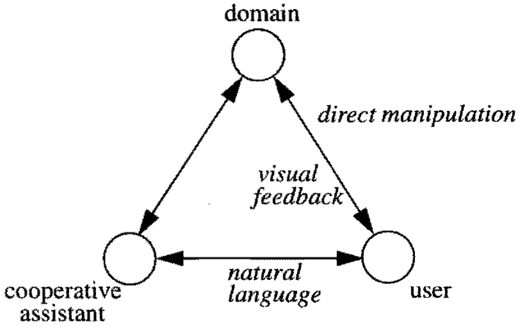


Figure 17 The DenK triangle of communication

## 8.4 The DenK triangle of communication

### 8.4.1 General aspects

The communicative situation in DenK can be represented in a triangle (Figure 17). The meaning of some of the terminology used in this triangle is similar to the meaning of some of the aspects of the systems that have been discussed before. A particular feature of this triangle, which is absent in the other systems, is that the DenK system is divided into two separate parts: the cooperative assistant and the domain of conversation (the *application domain*). Different modes of interaction are possible between the user and the cooperative assistant, the user and the domain, and the cooperative assistant and the domain.

Hutchins (1989) called this type of communicative situation, in which the user may interact with an intermediary that can act upon the world of action, or the user may act upon the world directly, the *collaborative manipulation* interface. This interface is a combination of the *conversation* interface, in which the user conducts a conversation by means of symbolic descriptions with an intermediary who acts on the world of action, and the *model-world* interface, in which the user takes action directly in the world of action which is itself the medium for the interface language (Hutchins, Hollan and Norman, 1986).

#### 8.4.1.1 The cooperative assistant

The cooperative assistant is the conversation partner of the user within the system (i.e., in Hutchins, Hollan and Norman's (1986) terms: the intermediary). The *knowledge representation* of the assistant is based on CTT (Constructive Type Theory) and is divided into private and mutual knowledge. Private knowledge can be, for instance, knowledge about features of objects in the domain that are not visible to the user. The mutual knowledge contains all the dialogue contributions and the actions that are carried out in the domain. The utterance that is generated by the assistant depends on the state of the domain, the dialogue and the different knowledge contexts.

The consequence of the cooperative assistant and the domain being separate is that domain knowledge is not necessarily represented in the assistant. Consequently, the assistant does not necessarily know all the features of the domain of conversation. This depends on the available modalities of communication to interact with the domain and on the previous knowledge about the domain that is available. To find out whether a certain proposition holds in the domain, the assistant may check the knowledge context or, if allowed, inspect the domain directly.

#### 8.4.1.2 The domain

The domain is seen as a representation of the 'real world' (i.e., in Hutchins, Hollan and Norman's (1986) terms: the model-world). Thus, it can be considered an *external information source*. The domain may contain graphical objects that may exhibit autonomous behavior. None of the other systems provide domain objects that can behave autonomously, i.e., that can alter their behavior independently of the user. In Table 9, the domain and the graphics are italicized to indicate that the domain and the objects represented in it are considered separate from the communication component (the cooperative assistant).

The current application domain is a simplified model of the electron microscope (which does not exhibit autonomous behavior). The electron microscope resembles the domains in WIP and CUBRICON in the sense that it provides a graphical representation of the actual domain that is the real device outside of the system. Users of these systems try to learn how to operate the real device, not just the graphical representation.

#### 8.4.1.3 Modes of interaction

The user, who is represented in the third angle in Figure 17, can communicate with the system in a variety of modes (*input modes*). Communication with the cooperative assistant takes place by means of natural language (English). The types of utterances that can be used are interrogatives and imperatives. It is neither possible (yet) to use indirect utterances, nor to use expressions concerning the system's knowledge (e.g., 'I think that...'). The reason is that the necessary semantics has not been worked out yet. Further, the user can observe the domain, point at graphical objects or manipulate them directly by means of the mouse.

The *output modes* of the system (the cooperative assistant) are natural language to communicate with the user (in this stage of the project only in the form of CTT expressions), and highlighting of objects to draw the attention to them. The graphics are not really considered output of the system, but independent features of the domain of conversation.

Only modalities are allowed that occur in real-life interaction as well. This means, for instance, that neither menu-mediated input (as in EDWARD and XTRA) nor Text buttons (as in AIMI) are possible in DenK, since these input modalities are not likely to occur in real life. Non-speech audio (as in AIMI) would only be possible as autonomous behavior of objects, or as a result of actions that are carried out in the domain, for instance, the sound of an object that is being dragged to another location.

### 8.4.2 Reference and focus of attention

The user can type in a natural language referring expression, and may choose to accompany this with a mouse-mediated pointing gesture. It is possible to refer to objects and to propositions. The latter type of reference is possible, because in CTT propositions are represented as types.

In the current version of the system the cooperative assistant can only produce CTT formulas. Only features that are part of the mutual knowledge can be used in referring expressions. Pointing at an object is done by highlighting it.

The user and the cooperative assistant can make a limited use of the spatial focus of attention through reduction of the referring expression. This is possible because the electron microscope domain contains some natural focus areas, such as a window (for instance, the window that contains the controls). It is not possible (yet) to make use of the functional focus of attention through reduction of the referring expression. The reason is that the system does not have a planning component, and the pre- and post-conditions of actions can not be determined yet.

## 8.5 Conclusions

None of the systems discussed here have the same type of communicative situation as the one in the DenK system. The most conspicuous difference is that they do not have a strict separation of the conversation partner and the application domain. This makes the systems less generic in the sense that the knowledge of the system about the domain can not be varied.

In DenK, the cooperative assistant forms the central part of the system. The contributions of the assistant to the interaction (either symbolic with the user, or physical with the domain) are knowledge-based. In addition, the assistant can be considered an expert in cooperative interaction. In all, this constitutes a natural situation in which two participants are communicating about a shared domain of conversation.

With respect to reference, many systems allow the use of natural language referring expressions, possibly combined with pointing gestures, both by the system and by the user. This way of referring seems to be close to the way people refer in natural real-world situations. This is the same way of referring that is advocated in the DenK communicative situation. However, DenK is the only system that considers mutual knowledge in choosing the contents of a deictic expression. A similar strategy is only followed in EDWARD, where the salience value of the target object is taken into account.

The spatial focus of attention is not included in most of the systems. The focus area is usually considered to contain all the graphics that are being shown on the screen. The focus is sometimes restricted by zooming in (ALFresco) or changed by a shift of the 'camera position' (COMET). In DenK, the window that is being talked about/manipulated is considered the focus area. There are plans to incorporate a more extended version of the spatial focus of attention (see Beun and Kievit, 1995).

DenK does not contain a notion of functional focus. Only CUBRICON and XTRA have mechanisms to find the referent with which the task can be most suitably carried out. Ambiguities in referential expressions are expected to be solved by using pointing gestures. However, since these systems are task-oriented, functional focus is very likely to occur, and provisions to deal with this type of focus (as well as spatial focus) would be a considerable contribution to attain cooperative (and natural) communication.

# 9 Conclusions and future research

## 9.1 Goal and method

The main goal of the research described in this dissertation was to get more insight in the factors that influence the way in which humans in a cooperative task-oriented dialogue refer (linguistically and gesturally) to objects that are physically present in a shared domain. The focus was on first references to objects, and in particular on the descriptive content, i.e., the part of the referring expression where features of the target object are expressed. Object reference was expected to be influenced by the particular communicative situation in which it is carried out.

Different communicative situations were modeled on the basis of the DenK triangle of communication. The communicative situation in the DenK system was used to ascertain that the results of the studies could contribute to the development of this generic graphical and natural language interface, in order to enhance the 'naturalness' of the interaction between the user and the communication partner internal to the system: the cooperative assistant.

Two empirical studies were carried out to investigate referring behavior in two different modalities of communication: spoken and via a computer keyboard. The domain that was chosen was a blocks domain in which a construction task was carried out by the participating subjects. Some of the most conspicuous empirical findings, which concerned the spatial focus of attention, were tested experimentally. All the results could be explained in the framework of a theory of minimal effort.

In order to interpret the findings in terms of the DenK system, an attempt was made to extrapolate the results from the blocks domain to DenK's current application domain: a model of the electron microscope. To get more insight in to how the generation and interpretation of object reference should be treated in DenK, and in particular, what the role of the focus of attention is in these processes, a number of similar systems was studied regarding these issues.

## 9.2 Conclusions

### 9.2.1 Main conclusion: Focus of attention

The main conclusion of the present study is that, both in spoken and in keyboard dialogues, the focus of attention plays a very important, perhaps until now underestimated, role in the generation and interpretation of object references. The term 'focus of attention' was used here to mean either the spatial focus of attention or the functional focus of attention. The spatial focus of attention represents a certain spatial area of the domain that is currently being attended to by the participants, usually the area closely surrounding the object that has been referred to previously, including the objects located within the area. The functional focus of attention is a subset of the objects in the domain that contains those objects with which the current action can be suitably carried out.

The most conspicuous finding is that the existence of either a spatial or a functional focus of attention can result in a reduction of the information included in the referring expression to refer to an object within the set of objects in focus in comparison with the amount of information that would be needed if the whole domain was to be taken into account. In the dialogues about the construction task in the blocks domain, the influence of the spatial focus of attention was found to be predominant.

#### 9.2.1.1 Spoken dialogues

The influence of the spatial focus of attention was most conspicuous in the spoken dialogues. Since speaking can be considered the most natural way of communication between humans, making use of spatial focus can be considered natural as well. The spatial focus was found to influence the referring behavior of the participants in at least three ways.

First, if possible, speakers prefer to select as the next target object one that is located within the current focus area.

Second, speakers use referring expressions to refer to objects within the focus area that would be ambiguous if the entire domain were to be considered, but that are minimally specifying if only the focus area is taken into account.

Finally, the location of the objects (either within or outside of the current focus area) influences the process of reaching mutual agreement between the participants that the reference has been understood and the object had been identified. It was found that it takes fewer turns to reach mutual agreement about objects within the focus area than about objects outside of the focus area. Also, a preference for using short noun phrases was found for objects within the focus area, and a preference for more complex noun phrases for objects outside of the focus area. Changes in the initial noun phrase were predominantly made (either by the initial speaker or by the addressee) at points where a transition to a new focus area took place.

#### 9.2.1.2 Keyboard dialogues

Keyboard dialogues between humans can be considered less natural than spoken dialogues, but they better resemble a common situation in human-computer interaction. The spatial focus of attention was found to play a role in these types of dialogues as well, although the effect was different from the spoken dialogues. An important problem that participants in keyboard dialogues encountered was the coordination of typing, gesturing, reading text on the computer screen, and inspecting the domain. This problem made it difficult to keep track of the current focus area.

A preference for choosing the next object to refer to within the current focus area did not occur in the keyboard dialogues. This could be explained by the observation that users found it difficult to type and inspect the domain at the same time in keyboard dialogues, which resulted in their being less able to keep track of the current focus area.

Further, the frequency with which speakers reduced the amount of information in referring expressions to refer to objects within the focus area was the same in the keyboard dialogues as in the spoken dialogues. In keyboard dialogues an additional reduction by means of gesturing occurred, which will be explained in section 9.2.2.

Finally, in the keyboard dialogues far more non-verbal (i.e., gestural) turns occurred than in the spoken dialogues. In both types of dialogues the mean number of turns (including the non-verbal turns) needed to establish mutual agreement was the same. However, in contrast to the spoken dialogues, no difference in the mean number of turns needed to refer to objects within and outside of the focus area was found. This result could again be ascribed to the problem of keeping track of which area constituted the current focus area.

#### 9.2.1.3 Experimental validation of spatial focus

It became clear in the two empirical studies that were performed concerning spoken and keyboard dialogues that the spatial focus of attention seemed to play a predominant role in the generation and interpretation of referring expressions. Since these results were not obtained under controlled conditions, at least two questions remained to be answered: (1) Do participants really consider a 'focus area' when interpreting a referring expression?, and (2) Is the fact that the expression is reduced essential for interpreting it as referring to an object within the assumed focus area?

To answer these questions an experiment was conducted to find evidence for the influence of an assumed focus area on the interpretation of referring expressions in three different grades of specification. And indeed, evidence was found for the existence of a focus area and its influence, which was different for different grades of specification, on the interpretation of referring expressions. If the grade of specification was low, subjects showed a preference for choosing the target object within the focus area. The higher the grade of specification, however, the larger the tendency to choose the target object in the non-focus area.

### 9.2.2 Theoretical framework: Minimal effort

#### 9.2.2.1 Focus of attention

Why participants make use of the spatial or functional focus of attention was explained by the observation that they adhere to the principle of minimal cooperative total effort. This principle states that humans cooperatively try to minimize the total effort expended to generate a referring expression, to interpret the expression and possibly ask questions for clarification, and to find the intended referent.

It was hypothesized that it takes less effort for both the speaker and the addressee to refer to an object that is in the current focus of attention. This is true for the speaker, since only the objects in focus have to be taken into account, and generally fewer words are needed. It is also true for the addressee, since a shorter expression generally takes less effort to interpret and the target object has to be searched for only within the set of objects in focus.

The effects of the spatial focus of attention, which were analyzed in more depth, were also formulated in terms of minimal effort. In the spoken dialogues, the effort was reduced by choosing as the next object to refer to an object within the current focus area, and by needing less information and fewer turns to refer to an object within the focus area. In the keyboard dialogues, the effort to refer to objects within the focus area was additionally reduced by using pointing gestures without any accompanying language. This sole pointing was probably used in keyboard dialogues because it was problematic (i.e., took more effort) to coordinate pointing and typing in a short demonstrative expression.

Spatial focus in relation to the grade of specification of referring expressions can also be explained in terms of minimal effort. If a low grade of specification is provided, it is assumed that this information is used to refer to the referent that requires the least effort to be found (i.e., within the focus area). If a higher grade of specification is given than is required to identify the referent within the focus area, this is more often interpreted as expressing a specific meaning (e.g., to switch to the non-focus area).

#### 9.2.2.2 Features

A second way in which the participants apparently tried to minimize effort was in the choice of features in the referring expression. In both spoken and keyboard dialogues participants preferred to use absolute features over relative features. Absolute features require less effort; for the speaker, since they generally require fewer words; for the addressee, since only the object having the features that were expressed should be taken into account. Relative features require more effort; for the speaker, since they generally require more words; for the addressee, since they require comparisons between objects in order to find the target object

### 9.2.3 Application of the results

On the basis of the results discussed above it was not *a priori* clear whether the scope of the importance of the focus of attention would be limited to the blocks domain only. In particular, it was not clear whether it would be of any relevance either in DenK's application domain or in other systems similar to the DenK system.

The application domain in the DenK system is a model of the electron microscope. Comparison of the electron microscope domain with the blocks domain shows that EM domain has fewer similar objects, the objects have names/labels, and objects are not located in spatial arrangements, but merely have meaningful hierarchical relationships.

It was concluded that reduction due to spatial focus is less likely to occur in this domain than in the blocks domain. The reason is that there are only a few types of objects (e.g., lenses) that occur more than once. Reference can be made to these objects by only mentioning the type (e.g., 'the lens') if a subpart of the domain where only one lens is located is considered as focus area.

In contrast, functional focus reduction is more likely to occur in the electron microscope domain, for instance, when giving commands to operate the controls. Since each control has one specific function it is usually sufficient to only utter the action that has to be carried out and not mention the object (control) that has to be operated.

None of the other multimedia systems that were studied employ a notion of spatial focus of attention in the way this has been discussed in this dissertation. This can be considered a deficiency in these systems. Most of the systems assume that users will use a combination of pointing and a demonstrative expression to refer to objects in the graphical domain. If either the referring expression or the pointing action is ambiguous, the system can usually solve the reference anyway by combining the two modalities. The DenK system could thus distinguish itself from other systems by incorporating an advanced notion of spatial focus of attention.

Functional focus, which has already been applied in some of the other systems, can not be implemented in DenK at present, since a planning mechanism has not been developed yet.

## 9.3 Limitations and future research

By discussing the results in terms of the DenK triangle of communication, it becomes clear that in the current investigation not all possible parameters that form part of this triangle have been studied. In other types of communicative situations, other types of object references and other ways to make use of the focus of attention are expected to apply. The principle of minimal cooperative total effort should be able to account for all these different uses.

### 9.3.1 The domain

Both the blocks domain and the electron microscope domain are *dynamic* in the sense that they can change, but only through actions carried out by the participants. In that respect, they differ from other types of dynamic domains, in which objects occur that can exhibit *autonomous behavior*, without interference by the participants. It would be interesting to see in what way a dynamically changing domain influences the way in which participants refer to the objects. For instance, this type of domain may induce the participants to refer to types of *movements* of objects. A particular movement pattern could then distinguish particular objects from other objects. A spatial focus of attention is less likely to occur in this type of domain, since the mutual spatial relationships between objects may change constantly. An example of an occurrence of a spatial focus area involving a moving object, is a situation where this object is constantly circling around a static object.

The domains that have been studied also differ from *static* domains, in which no changes can take place at all. In static domains no influence of functional focus is expected to occur, since no actions can be carried out in it.

The domains differ in the *types of objects* that were included in them. In the blocks domain the objects were quite similar, and more than one object of the same type occurred. The relations between the objects were mainly spatial. In contrast, in the electron microscope domain the objects clearly differed from each other, and their relations were mainly functional. Hence, the two domains seem to cover a wide range of the spectrum of all possible types of objects.

### 9.3.2 The dialogue participants

In the empirical studies the instructor and the builder fulfilled the roles of the user and the cooperative assistant, respectively. The knowledge was divided between these participants, such that the instructor knew the end result of the task to be carried out, and the builder experienced the physical limitations of actually constructing a block building in this particular domain. Both participants knew the features of the objects involved as well as the requirements of the task. The division of knowledge was known by both participants. The strategy that was applied to carry out the task as well as the roles of the participants in this task followed naturally from the division.

Other types of knowledge division between the user and the cooperative assistant are possible. In principle, unlike in comparable systems, in DenK every division of knowledge between the user and the assistant is possible, since the cooperative assistant and the domain are separate components. In the prototype of the DenK system, however, the user is a novice who wants to learn about the workings of an electron microscope, and the assistant is an expert on this particular piece of equipment. In this type of division the assistant knows about features of objects in the domain that are not known by the user.

The expert-novice communicative situation can have repercussions for the referring behavior. The assistant is restricted in the types of features to be included in referring expressions, since only features that are known to both participants are permitted.

### 9.3.3 The modalities of communication

The modalities of communication that were studied include the symbolic communication between the participants and the physical communication between each participant and the domain of communication.

With respect to the communication between the user and the cooperative assistant, both spoken and keyboard communication were studied. Other symbolic types of communication, such as facial expressions and bodily postures, were excluded from the studies. This type of interaction was not possible, since a screen was placed in between the participants. However, in human-computer interaction it is not really possible yet to interpret these types of interaction anyway.

Pointing gestures to refer to objects in the domain of conversation were also included in the studies. However, participants also used other types of gestures than merely pointing gestures, for instance, gestures to demarcate an area on the foundation plate and symbolic gestures like the ‘OK’ sign. These types of gestures were not included in the analysis. The former type of gesture would be interesting to study, though, since it contributes to the establishment of a spatial focus area, and accordingly, to the generation and interpretation of referring expressions.

Because of the presence of the screen between the participants, they could not observe each other's gaze direction. This can be seen as a deficit, because a gaze direction can act as an alternative for a pointing gesture. In human-computer interaction, it is possible to apply an eye tracking device to measure the gaze direction of users. This type of device would be very helpful in determining the focus area of the user, which is then assumed to be the area at which the gaze is directed. The system could use this information to interpret ambiguous referring expressions, by assuming that the target object is the object that both answers the description or is suited for carrying out the expressed action with and is located in the vicinity of the gaze direction of the user.

In the keyboard dialogues, it was found that coordination problems exist between typing, gesturing and inspecting the domain. If speech input were possible these problems would be solved. Perhaps the necessity of pointing could be abandoned if eye tracking devices were to be used, but this would probably not enhance the naturalness of the interaction.

The domain of conversation in the studies was a *shared* domain. This means that both participants had visual access to the domain, and that both were able to manipulate objects in the domain.[1] Situations in which the domain is not a shared domain are not so easy to imagine in a graphical interface. An essential characteristic of a graphical interface is that it makes the domain graphically observable to the user.

However, it is thinkable that users only have partial access to the domain. This may be the case if the domain is simply too large to fit on the screen (as is for instance the case in EDWARD) or if objects in the back are hidden by objects in the front. The partial access does not have to be the same for the user and the system. For instance, if the cooperative assistant is an expert about the domain, all knowledge about the domain including the objects that are not visible to the user is available to it. The system should be aware of this discrepancy and, for instance, should not use deictic expressions to refer to objects not visible to the user.

### 9.3.4 Some remarks on minimal effort

In this thesis the principle of minimal cooperative total effort was applied to explain the use of the descriptive content of first referring expressions to objects located in a shared task-oriented domain of conversation. Of course, the idea that humans try to minimize effort was not originally developed to explain only this type of behavior. In fact, when Zipf (1949) introduced the Principle of Least Effort he tried to apply it to human behavior in general, in the context of a science of 'human ecology'. We do not have to go that far, and can limit ourselves to human referring behavior. The Principle of Minimal Cooperative Total Effort can in principle be extended to explain other types of referring behavior and other parts of referential acts (i.e, determiners and gestures) than those that have been studied in this thesis.

*Other types of referential acts* include those used to introduce new objects in the

[1] Although in the empirical studies only the builder was allowed to manipulate the blocks, there was no physical reason why the instructor could not do the same. Only the requirements of the task prohibited this type of physical interaction.

domain, repeated, anaphoric reference to objects that are located in the domain, reference to objects that have been removed from the domain and may have been referred to before, and even non-physical objects that are part of the mutual knowledge of the participants. These types of references have already been discussed in chapter 2.

If other types of referential acts are integrated in a theory of minimal effort, different types of effort will also play a role. For instance, if the addressee encounters a referring expression, first a decision has to made whether it is being used deictically or anaphorically, and subsequently the referent has to be found in the physical domain or in the dialogue context, respectively.

The *determiners* that are used as part of the referring expression can guide the decision as to whether a referring expression is deictic or anaphoric (see Beun and Kievit, 1995). For instance, referring expressions including demonstrative determiners are most likely to refer to an object in the domain, whereas definite determiners usually occur in anaphoric references.

The type of demonstrative that is used in a deictic expression may also carry information about where to look for the target object within the physical domain. In a recent study by Piwek, Beun and Cremers (1995) that was based on the spoken dialogues it was found that proximates (*'deze'/'dit'*) ('this') were mainly used to refer to objects outside of the focus area, whereas distals (*'die'/'dat'*) ('that') were preferably used to refer to objects within the current focus area. Hence, it seems that demonstrative determiners help the addressee to direct his or her attention to the intended focus or non-focus area.

The use of referential *gestures* may also contribute to a minimization of effort. There are different types of gesture that contribute to the identification of a certain object, for instance, pointing gestures referring to an individual object, or gestures encircling a larger focus area. It is not *a priori* clear whether it is easier to use a gesture or a referring expression (spoken or written) in a certain situation. In the study by Piwek et al. (1995) on spoken dialogues it was found that pointing gestures were preferably used to refer to objects outside of the current focus area, i.e., to switch the attention to a new area. This makes sense in terms of minimal effort, since it usually takes a considerable number of words to switch to a new focus area, and it is probably easier to use a gesture instead.

Obviously, how to minimize different amounts of effort employed in different modes of interaction at a certain point in a dialogue remains a problem for a dialogue participant. In this thesis attempts have been made to unravel some of the ways in which this minimization can in principle be accomplished.

# References

Ahn, R. (1995) Logical model of the electron microscope. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, *DenK Report 95/15.*

Ahn, R.M.C., Beun, R.J., Borghuis, T., Bunt, H.C. and Overveld, C.W.A.M. van (1995) The DenK-architecture: a fundamental approach to user-interfaces. *Artificial Intelligence Review* 8, 431-445.

Allgayer, J., Harbusch, K., Kobsa, A. Reddig, C., Reithinger, N. and Schmauks, D. (1989) XTRA: a natural language access system to expert systems. *International Journal of Man-Machine Studies* 31, 161-195.

Alshawi, H. (1987) *Memory and context for language interpretation.* Cambridge, Cambridge University Press.

Ariel, M. (1990) *Accessing noun-phrase antecedents.* London, Routledge.

Bennis, H. and Hoekstra, T. (1983) *De syntaxis van het Nederlands: een inleiding in de regeer- en bindtheorie.* Dordrecht, Foris.

Beun, R-J. (1994a) De keuze voor het DenK-domein. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, *DenK Report 94/01.*

Beun, R-J. (1994b) Introduction to the electron microscope. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, *DenK Report 94/03.*

Beun, R-J. (1995) De electronenmicroscoop en de gevolgen voor de DenK-deelprojecten. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, *DenK Report 95/08.*

Beun, R.J. and Bunt, H.C. (1987) Investigating linguistic behaviour in information dialogues with a computer. *IPO Annual Progress Report* 22.

Beun, R-J. and Kievit, L. (1995) Resolving definite expressions in DenK. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, *DenK Report 95/16.*

Bos, E., Huls, C. and Claassen, W. (1994) EDWARD: full integration of language and action in a multimodal user interface. *International Journal of Human-Computer Studies* 40, 473-495.

Brachman, R.J. and Schmolze, J.G. (1985) An overview of the KL-ONE knowledge representation system. *Cognitive Science* 9, 171-216.

Brown, R. (1958) How shall a thing be called?. *Psychological Review* 65(1), 14-21.

Brown, G. and Yule, G. (1983) *Discourse analysis.* Cambridge, Cambridge University Press.

Bunt, H. (1994) Context and dialogue control. *Think Quarterly* 3, 19-30.

Bunt, H.C., Ahn, R., Beun, R-J., Borghuis, T. and Overveld, K. van (1995) Cooperative multimodal communication in the DenK project. In: Bunt, H., Beun, R-J. and Borghuis, T. (eds.) *Proceedings of the International Conference on Cooperative Multimodal Communication CMC/95, Eindhoven, May 24-26, 1995.* Tilburg/ Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, 115-128.

Burger, J.D. and Marshall, R.J. (1993) The application of natural language models to intelligent multimedia. In: Maybury, M.T. (ed.) *Intelligent multimedia interfaces.* Menlo Park, AAAI Press, 174-196.

Claassen, W. (1992) Generating referring expressions in a multimodal environment. In: R. Dale et al. (eds.) *Aspects of automated natural language generation.* Sixth international workshop on natural language generation, Trento, Italy, April 5-7, 1992.

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.

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

Cohen, P.R. (1984) The pragmatics of referring and the modality of communication. *Computational Linguistics* 10(2), 97-125.

Cohen, J.D., MacWhinney, B., Flatt, M. and Provost, J. (1993) PsyScope: a new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments and Computers* 25(2), 257-271.

Cremers. A. (1993) Referring behaviour in task dialogues. In: Sanford, A.J., Anderson, A.H., Moxey, L.M. and Gilhooly, K. (eds) *Abstract of papers.* Glasgow, International Conference on The Psychology of Language and Communication, University of Glasgow, 31st August - 3rd September 1993.

Cremers, A.H.M. (1994) Referring in a shared workspace. In: M.D. Brouwer-Janse and T.L. Harrington (eds.) *Human-machine communication for educational systems design (*NATO ASI Series, Subseries F, Computer and Systems Design 129). Berlin, Springer Verlag, pp. 71-78.

Cremers, A.H.M. (1995a) The process of cooperative object reference in a shared domain. In: Fava, E. (ed.) *Speech acts and linguistic research: proceedings of the workshop, July 15-17, 1994. A participant symposium held during the First International Summer Institute in Cognitive Science. Center for Cognitive Science, State University of New York at Buffalo, Buffalo, USA.* Padova, edizioni nemo, pp. 139-153.

Cremers, A.H.M. (1995b) Object reference during task-related terminal dialogues. In: Bunt, H., Beun, R-J. and Borghuis, T. (eds.) *Proceedings of the International Conference on Cooperative Multimodal Communication CMC/95, Eindhoven, May*

*24-26, 1995*. Tilburg/Eindhoven, SamenwerkingsOrgaan Brabantse Universiteiten, pp. 115-128.

Cremers, A. and Beun, R-J. (1995) Object reference in a shared domain of conversation. *IPO Manuscript 1089*. Submitted for publication in *Pragmatics and Cognition*.

Deutsch, W. and Pechmann, Th. (1982) Social interaction and the development of definite descriptions. *Cognition* 11, 159-184.

Deemter, K. van, Beun, R-J. and Vet, J. de (1995) Dialogue systems: taxonomy and architecture. Eindhoven, Institute for Perception Research, *Memorandum no. 430*.

Ekman, P. and Friezen, W. (1972) Hand movements. *Journal of Communication* 22, 353-374.

Feiner, S.K. and McKeown, K.R. (1990) Coördinating text and graphics in explanation generation. In: *Proceedings of the AAAI-90*, 442-449.

Glass, A.L. and Holyoak, K.J. (1986) *Cognition*. New York: Random House.

Goodman, B.A. (1986) Reference identification and reference identification failures. *Computational Linguistics* 12(4), 273-305.

Grice, H.P. (1975) Logic and conversation. In: P. Cole and J. Morgan (eds.) *Syntax and Semantics, Vol. 3: Speech acts*. New York, Academic Press, 41-58.

Grosz, B.J. (1977) *The representation and use of focus in dialogue understanding*. Technical Note 151. Menlo Park: SRI International.

Grosz, B.J. (1981) Focusing and description in natural language dialogues. In: A. Joshi, B. Webber and I. Sag (eds.) *Elements of discourse understanding*. New York: Cambridge University Press.

Grosz, B.J. and Sidner, C.L. (1986) Attention, intentions and the structure of discourse. *Computational Linguistics* 12(3), 175-204.

Hauptmann, A.G. and Rudnicky, A.I. (1988) Talking to computers: an empirical investigation. *International Journal of Man-Machine Studies* 28, 583-604.

Heim, I. (1982) *The semantics of definite and indefinite noun phrases*. (Dissertation). Amherst: University of Massachusetts.

Herrmann, Th. (1983) *Speech and situation: a psychological conception of situated speaking*. Berlin, Springer Verlag.

Huls, C., Bos, E. and Claassen, W. (1995) Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics* 21(1), 59-79.

Hutchins, E. (1989) Metaphors for interface design. In: Taylor, M.M., Néel, F. and Bouwhuis, D.G. (eds.) *The structure of multimodal dialogue*. Amsterdam, North-Holland, 11-28.

Hutchins, E., Hollan, J.D. and Norman, D.A. (1986) Direct manipulation interfaces. In: Norman, D.A. and Draper, S. (eds.) *User centered system design*. Hillsdale, Erlbaum.

Kirsner, R.S. (1979) Deixis in discourse: an exploratory quantitative study of the modern Dutch demonstrative adjectives. In: T. Givón (ed.) *Syntax and semantics, Vol. 12: Discourse and syntax.* New York, Academic Press.

Knapp, M.L. and Hall, J.A. (1992) *Nonverbal communication in human interaction.* Harcourt Brace Jovanovich College Publ.

Krauss, R.M. and Glucksberg, S. (1977) Social and non-social speech. *Scientific American* 236(2), 100-105.

Leeuwen, H. van (1993) The optical modes for TEM. *Philips Internal report* (Philips Electron Optics, Best)

Levelt, W.J.M. (1982) Linearization in describing spatial networks. In: S. Peters and E. Saarinen (eds.) *Processes, beliefs, and questions.* Dordrecht, Reidel.

Levelt, W.J.M. (1989) *Speaking: from intention to articulation.* Cambridge/London, The MIT Press.

Levinson, S.C. (1992) Primer for the field investigation of spatial description and conception. *Pragmatics* 2(1), 5-47.

Lewis, C. (1993) Analyzing means from repeated measures data. In: G. Keren and C. Lewis (eds.) *A handbook for data analysis in the behavioral sciences: statistical issues*. Hillsdale, Lawrence Erlbaum, 73-94.

Lewis, D. (1979) Scorekeeping in a language game. In: Bäuerle et al. (eds.) *Semantics from different points of view.* Berlin, Springer.

Loftus, G.R. and Mackworth, N.H. (1978) Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance* 4, 565-572.

Lyons, J. (1977) *Semantics* (2 vols.). Cambridge, Cambridge University Press.

Mackworth, N.H. and Morandi, A.J. (1967) The gaze selects informative details within pictures. *Perception and psychophysics* 2, 547-552.

Mason, J.A. and Edwards, J.L. (1988) Surveying projects on intelligent dialogue. *International Journal Man-Machine Studies* 28, 259-307.

McNeill, D. and Levy, E. (1982) Conceptual representations in language activity and gesture. In: R.J. Jarvella and W. Klein (eds.) *Speech, place and action: studies in deixis and related topics*. Chichester, Wiley and Sons, 61-71.

Olson, D.R. (1970) Language and thought: aspects of a cognitive theory of semantics. *Psychological Review* 77, 257-273.

Oviatt, S.L. and Cohen, P.R. (1991) Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language* 5, 297-326.

Neal, J.G. and Shapiro, S.S. (1991) Intelligent multi-media interface technology. In: J.W. Sullivan and S.W. Tyler (eds.) *Intelligent user interfaces*. New York, ACM Press, 11-43.

Pechmann, Th. (1984) *Überspezifizierung und Betonung in referentieller Kommunikation.* (Dissertation). Mannheim.

Piwek, P., Beun, R-J. and Cremers, A. (1995) Demonstratives in Dutch cooperative task dialogues. *IPO Manuscript 1134*, Submitted for publication in Computational Linguistics.

Piwek, P.L.A., Beun, R-J. and Cremers, A.H.M. (1996) Deictic use of Dutch demonstratives. *IPO Annual Progress Report* 30, 1995.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1972) *A grammar of contemporary English.* London, Longman.

Reithinger, N. (1987) Generating referring expressions and pointing gestures. In: G. Kempen (ed.) *Natural language generation: new results in artificial intelligence, psychology and linguistics.* NATO ASI Series E: Applied Sciences, no. 135. Dordrecht, Martinus Nijhoff, 71-81.

Schmauks, D. and Reithinger, N. (1988) Generating multi-modal output - conditions, advantages and problems. In: *Proceedings of the Twelfth international conference on computational linguistics, Budapest, Hungary*, 584-588.

Shapiro, S.C. (1979) The SNePS semantic network processing system. In: N.V. Findler (ed.) *Associative networks: the representation and use of knowledge by computers.* New York, Academic Press, 179-203.

Sperber, D. and Wilson, D. (1986) *Relevance: communication and cognition.* Cambridge, Harvard University Press.

Stock, O. (1991) Natural language and exploration of an information space: the ALFresco interactive system. In: *Proceedings of the Twelfth international joint conference on artificial intelligence (IJCAI'91)*, Sydney, 24-30 August. San Mateo, Kaufmann, 972-978.

Thomas, J.C. (1974) An analysis of behavior in the hobbits-orcs problem. *Cognitive Psychology* 6, 257-269.

Treisman, A.M. and Gelade, G. (1980) A feature-integration theory of perception. *Cognitive Psychology* 12, 97-136.

Wahlster, W., André, E., Finkler, W., Profitlich, H.-J., and Rist, Th. (1983) Plan-based integration of natural language and graphics generation. *Artificial Intelligence* 63, 387-427.

Wright, P. (1990) Using constraints and reference in task-oriented dialogue. *Journal of Semantics* 7, 65-79.

Zipf, G.K. (1949) Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, Addison-Wesley.

# Summary

This thesis reports a study on how humans refer to objects, when they participate in a dialogue that is concerned with a certain task they are carrying out cooperatively and in which these objects are involved. The study is part of the DenK program ('DenK' stands for 'Dialoogvoering en Kennisopbouw' in Dutch, which means 'Dialogue Management and Knowledge Acquisition'). In this program a generic interface is being developed, that allows the user to communicate with a system by means of both natural language and graphical interaction.

An important part of the communication via a linguistic and graphical interface consists of references to objects. On the one hand, the system should have knowledge about the way in which humans generate referring expressions, in order to be able to interpret the linguistic input of users. On the other hand, the system should be able to generate referring expressions in the same way as humans are used to do this, in order to make sure that the user experiences the dialogue as being as natural as possible. In order to study the referring behavior of humans in a relatively natural situation, an empirical set-up is designed that is analogous to the communicative situation within the DenK system. Pairs of subjects were asked to carry out a certain construction task in a blocks domain. In this task one of them played the role of instructor and the other the role of builder. Both spoken and typed (via computer terminals) dialogues were studied. Beside one of these two modalities of communication the participants were allowed to use (pointing) gestures too.

On the basis of the interaction that resulted from this task a model of referring behavior is developed that is formulated in the principle of minimal cooperative total effort. This principle says that both the speaker and the addressee try to spend as little effort as possible to, on the one hand, refer to an object, and, on the other hand, to interpret the reference and identify the target object. They accomplish this, among other things, by choosing the nature of the information in the reference such that it is easy for the hearer to do the correct identification. They also try to give exactly enough information in the reference, so that neither a question for clarification is necessary nor an incorrect identification takes place.

Speakers and typists appear to employ two important ways to minimize their cooperative effort: 1. by using absolute features in referring expressions, and 2. by reducing the information in referring expressions to objects within the focus area. By using absolute features, such as color (e.g., 'red'), speakers (and typists) ascertain that hearers (and readers) only have to search for objects having these particular features, and do not have to compare them to other objects. These advantages do not apply to relative features, for instance used when the location with respect to another object is being indicated (e.g., 'the red one next to the large blue one').

An important way to reduce the information in references is to make use of the so-called focus area. This is a sub-part of the domain at which the attention is focussed at that moment, for instance because an action has just been carried out there. Both to generate references to objects within this area and to identify them generally fewer objects have to be considered than when the whole domain is involved. This results in a shorter mean length of the references, and less effort to find the target objects. Also generally fewer turns are needed to reach mutual agreement that the target object has been identified.

Minimization of effort can also be achieved by making use of the phenomenon that a number of objects are in functional focus at a certain moment. These are objects that are most fit for carrying out the current task with. For instance, the blocks that are located at the top level of the block building can be removed most easily. If the task involves removal of a block, the referring expression used only has to distinguish the target object from the other top-level blocks.

The results with respect to spatial focus were tested in a more controlled experiment. Subjects were presented with a series of referring expressions in different grades of specification accompanied by a configuration of six blocks, on a computer monitor. They were asked to each time indicate the block they thought the expression was referring to. The presence of a focus area appeared to lead to a preference for choosing a block within this area. However, the higher the grade of specification of the referring expression, the more often subjects chose for a block in the non-focus area. Redundancy in referring expressions appears to be an indication for a transition to a new focus area.

Finally, the results of the studies were applied to another domain than the blocks domain, namely the domain of an electron microscope. This is the domain that has been chosen as the application domain of the DenK program. In this domain other types of objects occur that have other types of relationships. The focus area seems to play a smaller role here than in the blocks domain, most importantly because the objects in the electron microscope domain have names that can be used to refer to them. In the latter domain the function of objects also seems to play a larger role than their spatial ordering.

A comparative study of other natural language processing and graphical systems resulted in the observation that none of these systems contain a notion of focus area. In these systems the focus area is taken to be either the whole graphical image that is visible at that particular moment, or the window the user is working in at that moment. In these systems disambiguation of referring expressions usually takes place by connecting linguistic and graphical input in order to select a unique object. In order to enhance the naturalness of the interaction it is recommended to introduce the concept of focus area in such systems. In any case, the DenK system could distinguish itself by doing this.

# Samenvatting (Summary in Dutch)

Dit proefschrift vormt de neerslag van een onderzoek naar hoe mensen refereren naar objecten, als ze deelnemen aan een dialoog die betrekking heeft op een bepaalde taak die ze gezamenlijk uitvoeren en waarin deze objecten betrokken zijn. Het onderzoek maakt deel uit van het DenK-programma ('DenK' staat voor 'Dialoogvoering en Kennisopbouw'). In dit programma wordt een generiek interface ontwikkeld, dat de gebruiker in staat stelt om zowel met behulp van natuurlijke taal als door middel van grafische interactie met een systeem te communiceren.

Een belangrijk onderdeel van de communicatie binnen een talig en grafisch interface bestaat uit verwijzingen naar objecten. Enerzijds moet het systeem kennis hebben over hoe mensen referentiële expressies genereren om de talige invoer van gebruikers te kunnen interpreteren. Anderzijds moet het systeem ook zelf referentiële expressies kunnen genereren op de manier zoals mensen dat gewend zijn, om ervoor te zorgen dat de gebruiker de dialoog als zo natuurlijk mogelijk ervaart. Ten behoeve van de studie naar het verwijzingsgedrag van mensen in een betrekkelijk natuurlijke situatie, is een empirische opzet ontworpen die analoog is aan de communicatieve situatie binnen het DenK-systeem. Aan paren van proefpersonen werd gevraagd om een bepaalde constructietaak in een blokkendomein uit te voeren, waarbij één van beiden de rol van instructeur had en de ander de rol van bouwer. Zowel gesproken als getypte (via computer terminals) dialogen werden bestudeerd. De deelnemers mochten naast één van beide communicatiemodaliteiten ook (wijs-)gebaren gebruiken.

Op basis van de uit deze taak resulterende interactie is een gedragsmodel ontwikkeld dat is geformuleerd in het principe van minimale coöperatieve totale moeite. Dit principe stelt dat spreker en hoorder er naar streven om samen zo min mogelijk moeite te doen om enerzijds naar een object te verwijzen, en anderzijds de verwijzing te interpreteren en het object te identificeren. Dit bereiken ze onder andere door de aard van de informatie in de verwijzing zo te kiezen dat het voor de hoorder gemakkelijk is om de juiste identificatie te verrichten. Ook proberen ze in de verwijzing precies voldoende informatie te geven, zodat geen wedervraag naar meer informatie nodig is of een foute identificatie plaatsvindt.

Sprekers en typisten bleken twee belangrijke manieren te hebben om hun gezamenlijke moeite te beperken: 1. door absolute kenmerken in de verwijzingen te gebruiken, en 2. door in verwijzingen naar objecten binnen het aandachtsgebied de gebruikte informatie te reduceren. Door absolute kenmerken, zoals kleur (bijvoorbeeld 'rood'), te gebruiken om naar objecten te verwijzen bereiken sprekers (en typisten) dat hoorders (en lezers) slechts naar objecten met die bepaalde eigenschap hoeven te zoeken, en ze niet met andere objecten hoeven te vergelijken. Dit is wel het geval bij het gebruik van relatieve kenmerken, zoals wanneer de lokatie ten opzichte van een ander object wordt aan-gegeven (bijvoorbeeld 'de rode naast de grote blauwe').

Een belangrijke manier om de informatie in verwijzingen te beperken is door gebruik te maken van het zogenaamde aandachtsgebied. Dit is een deelgebied binnen het domein waar op dat moment de aandacht op gericht is, bijvoorbeeld omdat daar juist een bepaalde actie uitgevoerd is. Zowel bij de generatie van verwijzingen naar objecten binnen dit gebied als bij de identificatie ervan hoeven in het algemeen minder objecten in ogenschouw genomen te worden dan wanneer het gehele domein hierbij betrokken zou zijn. Dit resulteert in een gemiddeld kortere lengte van de verwijzingen, en minder moeite om het object te vinden. In het algemeen kost het ook minder beurtwisselingen om naar objecten binnen het aandachtsgebied te verwijzen.

Minimalisering van moeite kan ook bereikt worden door gebruik te maken van het verschijnsel dat een aantal objecten zich op een bepaald moment binnen de functionele aandacht bevinden. Dit zijn de objecten die het meest geschikt zijn om er de huidige taak mee uit te voeren, bijvoorbeeld de blokken die zich op de bovenste laag van het bouwwerk bevinden kunnen het gemakkelijkst verwijderd worden. In dit geval hoeven alleen deze blokken van elkaar onderscheiden te worden.

De bevindingen met betrekking tot het spatiële aandachtsgebied zijn vervolgens getest in een meer gecontroleerd experiment. Aan proefpersonen werd gevraagd om op basis van een serie verwijzende uitdrukkingen in verschillende informatiegraden die steeds samen met een zestal blokken verschenen op een beeldscherm, het blok aan te wijzen dat volgens hen bedoeld werd. Het bleek dat de aanwezigheid van een aandachtsgebied er toe leidde dat vaker voor blokken binnen dit gebied werd gekozen. Echter, naarmate de uitdrukkingen een hogere mate van redundantie hadden trad een verschuiving op naar de keuze voor een blok buiten het aandachtsgebied. Redundantie in verwijzingen blijkt een indicatie te zijn voor de overgang naar een nieuw aandachtsgebied.

Als laatste is getracht om de bevindingen van het onderzoek toe te passen op een ander domein dan het blokkendomein, namelijk dat van de electronenmicroscoop. Dit is het domein dat binnen het DenK-programma is gekozen als applicatiedomein. In dit domein komen andersoortige objecten voor die andersoortige relaties met elkaar hebben. Het aandachtsgebied lijkt binnen dit domein een minder grote rol te spelen dan binnen het blokkendomein, onder andere omdat in het electronenmicroscoop-domein de objecten namen hebben die gebruikt kunnen worden om er naar te verwijzen. Ook lijkt in het laatste domein de functie van objecten een grotere rol te spelen dan hun spatiële ordening.

Uit een vergelijkend onderzoek met andere natuurlijke taal-verwerkende en grafische systemen is gebleken dat geen van deze systemen een notie van focusgebied bevatten. In deze systemen wordt er vanuit gegaan dat het focusgebied de gehele grafische voorstelling is die op dat moment op het beeldscherm zichtbaar is, ofwel het venster waarbinnen op dat moment gewerkt wordt. Desambiguatie van referentiële expressies vindt in deze systemen vaak plaats door grafische en talige invoer aan elkaar te koppelen en zo een eenduidig object te selecteren. Om de natuurlijkheid van de interactie te verhogen verdient het aanbeveling om het concept focusgebied te introduceren in dergelijke systemen. Het DenK systeem zou zich hier in elk geval mee kunnen onderscheiden.

# Curriculum vitae

Anita Cremers was born in Eindhoven on November 26, 1960. In 1979 she obtained the Atheneum-B certificate from the Anton van Duinkerkencollege in Veldhoven. In the same year she started the 3-year HBO education for 'Bibliothecaris-Documentalist' (librarian-documentalist) at the Bibliotheek- en Documentatie Academie in Tilburg. In 1982 she graduated, and started working at Shell International Petroleum Company in The Hague, first as a librarian, and later as a documentalist. In 1986 she quit her job and started studying Language and Literary Studies with a specialization in computational linguistics at the Katholieke Universiteit Brabant in Tilburg, and graduated cum laude in 1991. During this period she also worked, both as a student-assistant at the faculty and as a part-time librarian at the Openbare Bibliotheek (public library) in Tilburg. In August 1991 she became a researcher at the Institute for Perception Research (IPO) in Eindhoven. Her project was called 'Experimenteel onderzoek in mens-machine interactie' (Experimental research in human-machine interaction), and formed part of the DenK (Dialogue Management and Knowledge Acquisition) project, a cooperation between the Katholieke Universiteit Brabant, the Technische Universiteit Eindhoven, and IPO. This thesis is the result of that project.

# Stellingen

behorende bij het proefschrift
*Reference to objects: an empirically based study of task-oriented dialogues*
van A.H.M. Cremers

I. Deelnemers aan taakgerichte dialogen maken gebruik van het aandachtsgebied bij het formuleren van referentiële expressies naar objecten die zich binnen dit gebied bevinden. De toegestane communicatiemodaliteiten hebben echter invloed op de mate waarin van het aandachtsgebied gebruik gemaakt wordt. In gesproken dialogen komt dit bijvoorbeeld vaker voor dan in getypte dialogen. Het is daarom maar de vraag of in getypte mens-machine dialogen het aandachtsgebied een prominente rol speelt.

Dit proefschrift, hoofdstuk 5.

II. Redundantie in taal, met name in referentiële expressies, kan niet zonder meer verklaard worden als een gevolg van een productieprobleem of als het bewust geven van meer informatie om de hoorder te helpen de uiting te interpreteren binnen het huidige aandachtsgebied. Expressies die redundant zijn binnen het aandachtsgebied kunnen ook een indicatie zijn voor de hoorder om de referent buiten dit gebied te zoeken.

Dit proefschrift, hoofdstuk 6.

III. Bij het gebruik van absolute kenmerken in referentiële expressies (bijvoorbeeld 'de rode') hoeft slechts één object geïdentificeerd te worden, terwijl dat er bij relatieve kenmerken (bijvoorbeeld 'de rode naast de gele') minstens twee zijn. Om de moeite te minimaliseren die het genereren en interpreteren van object referenties met zich meebrengt, zal veelal de voorkeur worden gegeven aan absolute kenmerken. Dit is met name effectief in een type gespreksdomein waarin zich een grote variëteit aan objecten bevindt.

Dit proefschrift, hoofdstuk 3.

IV. Uitspraken waarin zeer nadrukkelijk één of meer Maximen van Grice worden geschonden, bijvoorbeeld bij het gebruik van referentiële expressies, zoals het geval is in het fragment dat is afgedrukt in het voorwerk van dit proefschrift, worden in het algemeen door toehoorders als humoristisch ervaren.

H.P. Grice (1975) Logic and conversation. In: P. Cole and J.L. Morgan (eds.) *Syntax and semantics 3: Speech acts.* New York, Academic Press, p. 41-58.

V. De gewoonte om in wetenschappelijke literatuur handelend over mens-mens dialogen onderscheid te maken tussen de spreker en de hoorder door naar de eerste met 'hij/hem' en de tweede met 'zij/haar' te refereren is ongewenst, omdat hierdoor de traditionele rolverdeling tussen mannen en vrouwen wordt benadrukt.

Deborah Tannen (1991) *Je begrijpt me gewoon niet: hoe vrouwen en mannen met elkaar praten.* Amsterdam, Prometheus. Vertaling van: *You just don't understand* (1990).

VI. Voorzover de huidige ontwikkelingen binnen de (commerciële) media in Nederland door veel kijkers/luisteraars als zeer belangwekkend worden gezien, is dit hoofdzakelijk een gevolg van het feit dat diezelfde media in hun eigen nieuwsbulletins onevenredig veel zendtijd aan deze ontwikkelingen besteden.

VII. De toename van de informatie die ons bereikt en de mogelijkheid om steeds meer informatie te raadplegen (bijvoorbeeld via Internet) leidt in vele gevallen niet tot het nemen van beter gefundeerde beslissingen. Deze situatie zal leiden tot ofwel een grotere mate van besluiteloosheid, ofwel een meer intuïtieve manier van besluitvorming.

VIII. De verschuiving van een tekstcultuur naar een beeldcultuur, die in de huidige maatschappij plaatsvindt, kan gezien worden als een betreurenswaardige terugkeer in de evolutie, maar ook als een verheugende rehabilitatie van de visuele 'taal'. Met name moderne beeldende kunstenaars kunnen van deze ontwikkeling profiteren.

IX. De veel gehoorde klacht van romanschrijvers dat zij door hun lezers onterecht worden vereenzelvigd met de hoofdpersonen in hun romans, is een moeilijk te vermijden artefact van het schrijven van fictie. Het doel en het plezier van het lezen van fictie ligt in het ervaren van gebeurtenissen door de ogen van iemand anders. Om deze ervaring op te wekken moeten schrijvers een sterk subjectieve taal bezigen, bijvoorbeeld door het gebruik van de 'ik-vorm'. Hierdoor wordt de suggestie gewekt dat schrijver en protagonist dezelfde persoon zijn.

Lynne E. Hewitt (1995) Anaphor in subjective contexts in narrative fiction. In: Judith F. Duchan, Gail A. Bruder and Lynne E. Hewitt (eds.) *Deixis in narrative.* Hillsdale, Lawrence Erlbaum Associates, p. 325-339.