Multimodal Reference to Objects: An Empirical Approach

Robbert-Jan Beun[1] and Anita H.M. Cremers*[2]

[1] Institute for Information and Computing Science,
Utrecht University, Utrecht, The Netherlands
rj@cs.uu.nl
[2] TNO Human Factors, Soesterberg, The Netherlands.
cremers@tm.tno.nl

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 multimodal 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 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 empirical study 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.

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* with multimodal access, i.e., a domain to which both participants have visual as well as physical access, they can communicate these objects by means of *referential acts*, i.e., verbal referential expressions and/or nonverbal references, such as pointing or other gestures. In an actual interactive situation, the speaker may use one or more of the object's features to indicate that 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', possibly in combination with a pointing action.

* This chapter is a slightly adapted version of: Beun, R.J. and Cremers, A.H.M. (1998) Object Reference in a Shared Domain of Conversation. Pragmatics and Cognition 6(1/2), 121–152.

The primary goal of this chapter is to present some fundamental cognitive concepts and pragmatic principles in 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 effort* (Clark and Wilkes-Gibbs, 1986); we will not only be concerned with the minimization of the effort to verbalize the expressions in a conversation, but also with 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. First, 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. Second, we 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. Since we are especially interested in the amalgam of the processes of object identification and object reference, we will restrict our analysis to first references to target objects. In these cases the descriptive content contains the maximal amount of information and the salience of the objects is not predominantly determined by the discourse.

To find evidence in real discourse for the hypotheses that we formulated on the basis of the principle of minimal cooperative 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 2, we define referential acts, focusing on the descriptive content of these acts. In section 3, we introduce the principle of minimal cooperative effort, which we think is the basic underlying mechanism for object reference, and discuss two important notions: 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 4, the setup of the empirical study that was carried out is described. Section 5 links the abstract notions used in the model to the properties of the domain used in the empirical study. In section 6, the results are described and in section 7 we will discuss how these results can be interpreted in terms of the model sketched in section 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 dagger †.[3]

referential act = *(referential expression) (†)*
referential expression = *(determiner) (descriptive content)*
descriptive content = *(premodifier)* (head) (postmodifier)**

Examples of referential acts are:

1. $(\text{het})_{det}((\text{grote})_{premod}(\text{rode})_{premod}(\text{blok})_{head}(\text{voor mij})_{postmod})_{descr.cont}$
   'the large red block in front of me'
2. $(\text{een})_{det}((\text{groot})_{premod}\ (\text{blok})_{head}\ (\text{dat achter de rode staat})_{postmod})_{descr.cont}$
   'a large block lying behind the red one'
3. $(\text{die})_{det}\ ((\text{grote})_{premod}\ (\text{hier})_{postmod})_{descr.cont}\ (\dagger)$
   'that large one here (†)'

The 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.

2.1 Descriptive content

The descriptive content 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

[3] Actually, in English as well as in Dutch, the form of references can be more complicated (Quirk et al., 1972; Bennis & Hoekstra, 1983). A reference may be constructed of: *(predeterminer)(determiner)(postdeterminer)*(premodifier)* (head)(postmodifier)** or *pronoun*. However, we will only consider the simple form here. Moreover, although reference to objects can also be carried out by using *proper names*, such as 'De Nachtwacht' ('The Nightwatch'), 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 name and object, and no alternative objects need to be considered for identification.

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 Pragmatic principles and cognitive concepts in object reference

3.1 The principle of minimal cooperative effort

Our analysis will be based on the *principle of minimal cooperative effort* defined by Clark and Wilkes-Gibbs (1986). They state that reference to objects can be seen as a collaborative process; the principle 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.[4]

In this chapter, we will assume that the principle should be interpreted in a broad sense. In contrast to Clark and Wilkes-Gibbs, who focus on the linguistic and dialogue-related aspects of the referential process, we emphasize the process of identification by the hearer. 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 and,

[4] 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.

as a result, try to minimize 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, either implicitly or explicitly. He can do this by making use of absolute features that require the identification of only one object.

3.2 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.

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, if it is related in some way 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. Below, we will briefly discuss the first two types of focus.

Inherent Salience Objects that are salient within the domain of conversation attract attention.[5] 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

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

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 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 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.'*

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. Since we concentrate in this chapter on first reference to objects, we will mainly consider focus in the domain of conversation. However, the focus of attention within the dialogue often coexists with a focus of attention within the domain of conversation.

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.

Beside the inherent salience of objects that may attract attention, which was discussed in the previous subsection, 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 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.'*

3.3 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 only one of the predominant criteria for choosing a particular feature. Speakers also have the choice to use either absolute or relative features to refer to a certain object. From the principle of minimal 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: 'the block that is located at the coordinates 318, 248' and '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 3: *'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 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 4: *'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.

3.4 Reduced Information

In the results and the discussion below, we will express reduction of information in terms of ambiguity and redundancy of the referential act with respect to a competitive set of objects. We will say that a referential act is ambiguous if two or more objects fit the description of the act; the act is redundant if any part of the descriptive content can be left out without becoming ambiguous. A referential act that is neither redundant, nor ambiguous will be called optimal.

The notions of ambiguity and redundancy will be applied to the current focus area as well as the whole domain (see figures 1 and 2). For example, in

a domain with two yellow blocks and a blue block, of which one yellow block and the blue block are present in the current focus area, the expression 'the yellow one' is ambiguous with respect to the whole domain, but optimal with respect to the focus area. We also include the pointing act of the speaker in our definition; so, if the speaker in the previous example also would have pointed to the yellow block, the referential act would be redundant with respect to the focus area and the whole domain, but not ambiguous. Unambiguous pointing actions combined with descriptive features (e.g., 'the yellow one †') are always considered as redundant.

Notice that, if a focus area is present, ambiguity of the referential act within the focus area always implies ambiguity within the whole domain. Vice versa, redundancy within the domain always implies redundancy within the focus area. Also, the definition implies that the referential act can never be both ambiguous and redundant with respect to the whole domain.

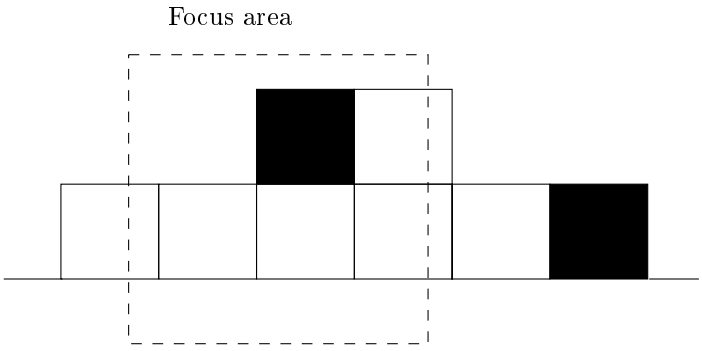

Fig. 1. The utterance 'the black one' is redundant with respect to the current focus area, but optimal in the domain

4 Empirical Setup

In order to find evidence for the hypotheses that were formulated in the previous section, we carried out an empirical study during which two participants were asked to perform a specific task in a shared domain of conversation. The situation is depicted in Figure 3 and can be described as follows.

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,

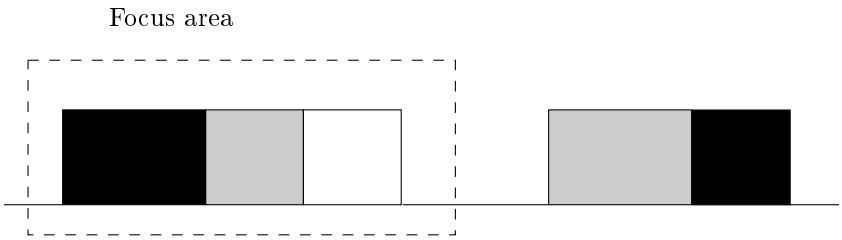

Fig. 2. The utterance 'the big black one' is redundant with respect to the current focus area, but optimal in the domain

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.

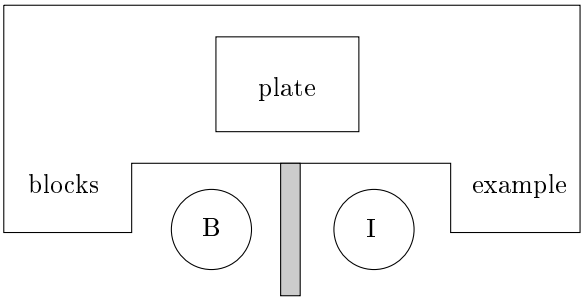

Fig. 3. Experimental configuration (top view), B=Builder, I=Instructor

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).[6] Schematic pictures of the 29 blocks that were involved in the building sessions are provided in Figure 4. 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.

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

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

10 building sessions were recorded on video-tape and the spoken communication was transcribed (Cremers, 1993). The dialogues that occurred during the sessions were similar to Grosz's task dialogues (Grosz, 1977).

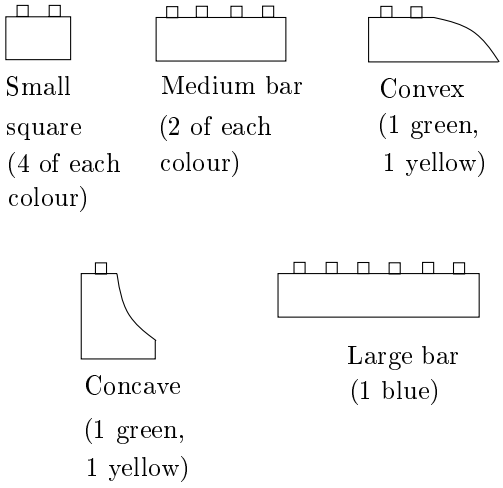

Fig. 4. Types, numbers and colours of blocks used in the experiment (side view)

5 Domain Properties and Definitions

Before we discuss the results in terms of the model sketched in Section 3, we first have to convert the abstract notions such as 'salience', and 'focus' into concrete domain properties.

5.1 Inherent Salience

During the rebuilding task, blocks were removed and others were added. On the average, 24 blocks were present on the foundation plate, only two of them were convex or concave. So, due to their deviated form and their relative small number of occurrence, these two types were considered inherently salient with respect to the bar and square types.

Although it can be argued that, due to the perceptual properties of the eye, yellow objects are inherently more salient than other coloured objects, we did not include this in our analysis. This was done because the colour feature was randomly distributed over the objects and yellow appeared almost equally often as the other colours.

5.2 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. We have distinguished five indicators that determine if an object is located in the current focus area. Occurring indicators are either domain-related or linguistic criteria, or combinations of both types.[7]

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 are used indicating to stay at the same location or that the (sub-)task has not yet been finished (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.[8] 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. (Dialogue 10.21-22; Cremers, 1993)

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 die rode op de grond staan.`
B: `Ja ja.` (haalt grote kleine gele, kleine blauwe eraf)

[7] In the list of criteria no task-related indicators are added. The possibility exists that the addressee is aware 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-related effects on the choice of references have been treated in depth by Grosz (1977).

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

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 the red one stays on the ground.*
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 current domain, since at least one more red block was located at the foundation plate. Also, the definite expression 'die' ('the') is used. Furthermore, the uses of 'blijft' ('stays') and 'alleen' ('only') 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.

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
- 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. (Dialogue 2.63-64; Cremers, 1993)

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 een groen`
`blokje.`

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 a green block.*

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').

6 Results

6.1 General Observations

During the execution of the task that was explained in section 4, the subjects used a total of 665 referential acts. Of these references, 145 were first references to objects located in the domain of conversation. Below we consider only these 145 first references.

Spatial focus: Based on the criteria formulated in the previous section, we were able to identify 45 objects (31%) out of the spatial focus area as a result of a focus change and 100 objects (69%) in the focus area at the time of the utterance; these results were scored independently by the two authors. In only four cases we had initial disagreement, but we decided on the criterion of linguistic markers that were used to move to another location (2 cases) or that indicated that the previous task had been finished (2 cases).

Pointing: In 69 cases (48%) a pointing act was used. The total number of pointing acts is slightly biased though, because three subjects declared afterwards that they tried to carry out the task without pointing. (The subjects were only told that they were allowed to point, not that they had to.) Leaving out these three subjects, the percentage of pointing actions was 68%.

Ambiguity and redundancy: Ambiguous references with respect to the whole domain occurred in 62 cases (43%); if a focus area was present, ambiguity with respect to the current focus area occurred in 12 cases (12%; 12 out of 100). In 44 cases (30%) the expression was redundant with respect to the whole domain; if a focus area was present, in 28 cases (28%; 28 out of 100) the expression was redundant with respect to the focus area. In 39 cases (27%) the reference was optimal.

Salience: In total, 13 references to salient objects were counted, such as 'the green slide' (concave type) or 'the half rounded one' (convex type). In 3 of these cases the referential act was redundant with respect to the whole domain, but it was never redundant in cases where the salient object was inside the focus area (3 times). Only 2 expressions were clearly ambiguous with respect to the whole domain.

Descriptive features: Of the total amount of 145 first referential acts that occurred in the dialogues, 90 (62%) just included absolute features (colour and/or shape). In 2 (1%) of the cases only relative features were used. In 27 cases (19%) combinations of relative and absolute features were used. Beside relative and absolute features, demonstrative expressions accompanied by a pointing action were used in 26 cases (18%).

Explicit relata: In only 19 cases (13%) an explicit relatum was used in the referential expression. In none of these cases a pointing action was used. Sometimes, the relatum refered to an object that was not a block (e.g., 'the floor', the participants), or was an abstract object (e.g., 'the second level'). In 4 cases one or two of the participants were mentioned as relatum, in 13 cases some object in the domain served as an explicit relatum, and in 2 cases both a participant and an object were explicit relata. If a domain object was used as a relatum (15 cases), in 10 cases this object was mentioned previously. In one case the relatum was located in the current focus area. In the 4 remaining cases the relatum was either inherently salient or a unique object within the domain.

6.2 The Influence of Changing the Focus Area

In Table 1 we have indicated a. the number of pointing actions, b. the salience of the target object, and c. the ambiguity and d. redundancy of the referential act with respect to the domain as a function of in or out focus of the target object. '+' indicates that these characteristics are present; '-' indicates that they were absent. For instance, in 10 cases where the referential act referred to a salient object, the object was out of focus; in 97 cases the object was in focus, but not salient.

Table 1. The number of pointing actions, the salience of the target object, and the ambiguity and redundancy of the referential act with respect to the domain as a function of in or out focus of the target object. '+' indicates that these characteristics are present; '-' indicates that they were absent.

	Pointing		Salient object		Dom. ambiguity		Dom. redundancy		Total
	+	-	+	-	+	-	+	-	
Out Focus	25	18	10	35	10	35	23	22	45
In Focus	44	58	3	97	52	48	21	79	100
Total	69	76	13	132	62	83	44	101	145

Except for the pointing action, the differences between in or out focus results differed significantly (Pointing: $\chi^2_{df=1} = 2.34$, $p < 0.2$; Salient object:

$\chi^2_{df=1} = 14.7$, $p < 0.001$; Domain ambiguity: $\chi^2_{df=1} = 10.6$, $p < 0.005$; Domain redundancy: $\chi^2_{df=1} = 13.28$, $p < 0.001$). In other words, redundancy of the referential expression appeared relatively more often when the target object was out of focus; vice versa, ambiguity appeared relatively more often when the object was in focus. Moreover, when a focus change appeared, relatively more reference was made to a salient object.

We did not find significant differences in the use of particular descriptive features (e.g., colour, shape or the use of relata) as a function of being in or out focus of the target object.

7 Discussion

We will now discuss the outcome of the empirical study in more detail and relate the results to the hypotheses discussed in Section 3.

7.1 Salience

Our first hypothesis was that *if the object is inherently salient within the domain of conversation, the speaker uses reduced information.*

In other words, we would expect most of the referential acts to the concave and convex objects ambiguous with respect to the whole domain. As we can read from the results, however, only two of the 13 references were ambiguous, and even less expected, 3 cases were redundant. These numbers are relatively low, so we should be careful to draw too many conclusions from this.

But let us go a little more deeply in these redundant cases. In two cases, redundancy was caused by the combination of a pointing act and a descriptive feature in the referential act; in one case, the redundancy was caused by the appearance of two descriptive features 'green' and 'slide', instead of 'slide' only. So the redundancy is at least minimal, caused by only one extra descriptive feature.

Also, in all three cases the redundancy appeared when the target object was not in the current focus area. Important here are the results of the redundancy of descriptions in general. As we can see in Table 1, in general descriptions of objects out of focus contain significantly more redundancy than descriptions of objects in focus. But the descriptions of salient objects out of focus contain redundancy in only 33% of the cases (3 out of 10), while descriptions of non-salient objects out of focus contain redundancy in 57% of the cases (20 out of 35). Due to the low total number of salient object descriptions the difference is not significant ($\chi^2_{df=1} = 1.87$, $0.1 < p < 0.2$), but there is at least a strong tendency for reduction of descriptive features when the object is salient.

So, the result is in line with the hypothesis and we would expect a strong tendency for ambiguity if more salient objects would be in focus. Testing this in a natural dialogue situation will often be difficult, though, since salient objects are always limited in number and are often picked out as a marker for establishing a new focus area (see Table 1). An important conclusion is that reduction of

information cannot simply be explained in terms of ambiguity or redundancy with respect to the whole domain, but that at least a distinction has to be made between objects in focus and objects out of focus.

7.2 Redundancy of information

From the second hypothesis, i.e., *'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'*, we would expect that most of the references are ambiguous with respect to the domain or at least not redundant with respect to the focus area. This was indeed supported by the results. In 28%, however, we still noticed redundancy in the act, but this redundancy was always caused by an extra pointing act, not by the addition of extra descriptive features.

In only three of the 52 cases the ambiguity caused an identification problem for the hearer, so both speaker and hearer not only used a focus area to reduce the information, but must have been mutually aware about each others focus area.

From the third hypothesis, i.e., *'use only information that distinguishes the target object from other objects that would also be suitable for use in carrying out the current action'*, we would expect that some of the references are ambiguous within the focus area and can be resolved by means of functional information. Apart from one case where the ambiguity could not be resolved without extra dialogue acts, in all other cases (11) the resolution process was supported by functional information. 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.

Although these results strongly support the hypotheses on redundancy of information, again we have to be careful to draw too many conclusions here. First, sometimes the unequivocal determination of a specific focus area is difficult and it may be the case that sometimes specific objects were in the focus area without being classified as such and vice versa. Therefore, to determine the focus area, we only included those objects where the objects were adjacent to the target object.[9] Second, in some cases spatial and functional information cannot easily be distinguished, since objects in the neighbourhood are often both functionally relevant and in the focus area.

An important finding is that references to objects outside the focus area are significantly more redundant than references inside the focus area. A reason for this redundancy is problably that the speaker is simply unable to overview the whole domain in short time and, therefore, cannot decide which and how many of the possible features to use in order to minimize the contribution (see also Pechmann, 1984). Probably 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

[9] Note that in these cases we did not have linguistic information at our disposal.

explained by realizing that speakers probably give more information to avoid an explanatory sub-dialogue in case the hearer has not understood the initial expression. So, the principle of minimality is still maintained, but not on the level of descriptive features, but on the level of identification and speech act turns.

7.3 Descriptive Features

The third hypothesis, *'use absolute features as much as possible and use relative features only if necessary'* is strongly supported by the data. Only 20% of the references contained relative features. But again we should be careful in our conclusions, since these numbers may highly depend on the domain, its properties, the task and the communicative situation. It may well be that in other situations, for instance, where pointing is impossible, or where objects are significantly different and absolute features play just a minor role (e.g., 'Look at the man with the funny hat'), the data on referential acts may not support the hypothesis in such a convincing way.

Finally, the fourth hypothesis, *'if an explicit relatum is needed for referring to the target object, choose as relatum an object that is in the focus of attention'*, is also strongly supported by the data. Either participants or other objects or both were used as 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.

Domain 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), are inherently salient or a unique object. 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. As can be seen in the results, in all cases the relata fulfilled these requirements.

7.4 Focus and the principle of connectivity

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 (69% in focus, 31% out of focus). In terms of the principle of minimal cooperative 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 the principle of connectivity, which predicts that "a speaker will go over a pattern as much as possible "without lifting the pencil", the mental pencil's point being the speaker's focus of attention" (Levelt, 1982: p.140). In Levelt's case, 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 process probably works in the same way as the 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 of attention 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'.

7.5 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. For instance, based on the same Dutch data, it has been shown in Piwek, Beun & Cremers (1995) that proximate demonstratives are used in cases were the speaker wants to signal to the addressee a need for extra effort to find the intended referent, while distals are used in cases where the referent is more 'given' with regard to the addressee's consciousness.

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, 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 the previous subsection and illustrated in example (9).

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.

8 Conclusions

In this chapter, we have tried to describe the 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 done so in line with Clark and Wilkes-Gibbs' principle of minimal cooperative effort and payed especially attention to the amalgam of the processes of object identification and object reference. From the principle 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 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.

By means of an empirical study, we were able to show that focus is not only a discourse-related phenomenon, but also a result of particular properties of the domain of conversation combined with the perceptual abilities of the dialogue partners. 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 more than half 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.

An important finding from the experimental study was that references to objects outside the focus area are significantly more redundant than references inside the focus area. This showed that the notion of reduction of information is a complex matter that cannot simply be explained in terms of redundancy or ambiguity of information with respect to the whole domain. With respect to the type of descriptive features, we did not find significant differences between in or out focus references.

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, and other modalities of communication. Since the concepts introduced in this chapter are basic properties of almost every human communication situation, we expect, however, the results to be relevant for a broad field

of applications.

Acknowledgements

This research was funded by the Universities of Brabant Joint Research Organization (SOBU). We would like to thank Kees van Deemter and Paul Piwek for extensive and useful comments on earlier drafts of this chapter.

References

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.

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

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.

Cremers, A.H.M. (1993) *Transcripties dialogen blokken-experiment (Transcriptions dialogues blocks-experiment).* IPO Report no. 889. Eindhoven: Institute for Perception Research.

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).* Heidelberg: Springer Verlag, 71–78.

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.

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

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

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

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

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 and London: The MIT Press.

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.

Mackworth, N.H. and Morandi, A.J. (1967) The gaze selects informative details within pictures. *Perception and psychophysics* 2:547–552.

Olson, D.R. (1970) Language and thought: aspects of a cognitive theory of semantics. *Psychological Review* 77:257–273.

Pechmann, Th. (1984) *Überspezifizierung und Betonung in referentieller Kommunikation.* (Dissertation). Mannheim.

Piwek, P.L.A., Beun, R-J. and Cremers, A.H.M. (1995) Deictic use of Dutch demonstratives. *IPO Annual Progress Report, 30.* Eindhoven: Institute for Perception Research.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1972) *A grammar of contemporary English.* London: Longman.

Sperber, D. and Wilson, D. (1986) *Relevance: communication and cognition.* Cambridge: Harvard University Press.

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.

This article was processed using the LaTeX macro package with LLNCS style