Running head: GENERATING MULTIMODAL REFERENCES

Generating Multimodal References

Ielka van der Sluis
Computing Science
University of Aberdeen

Emiel Krahmer
Communication & Cognition
Tilburg University

Abstract

This paper presents a new computational model for the generation of multimodal referring expressions, based on observations in human communication. The algorithm is an extension of the graph-based algorithm proposed by Krahmer et al. (2003) and makes use of a so-called Flashlight Model for pointing. The Flashlight Model accounts for various types of pointing gestures of different precisions. Based on a notion of effort the algorithm produces referring expressions combining language and pointing gestures. The algorithm is evaluated using two production experiments, with which spontaneous data is gathered on controlled input. The output of the algorithm coincides to a large extent with the utterances of the participants. However, an important difference is that the participants tend to produce overspecified referring expressions while the algorithm generates minimal ones. We briefly discuss ways to generate overspecified multimodal references.

Introduction

Human-computer interaction (HCI) studies the interaction between human users and computers which takes place at the user interface. Advances in HCI provide evidence that the use of multiple modalities, such as speech and gesture, for both input and output may result in systems that are more natural and efficient to use (Oviatt 1999). Consequently, current research in HCI shows an increased interest in developing interfaces that closely mimic human-human communication, and the development of "virtual characters" or "embodied conversational agents" (ECAs) that are able to communicate both verbally and non-verbally about a concrete spatial domain clearly fits this interest (e.g. Kopp et al., 2003; Cassell et al., 2000). A subtask that is addressed in many systems is that of identifying a certain object in a visual context accessible to both user and system. This can be done for example by an ECA that points to the object, possibly in combination with a linguistic referring expression (RE). With the design of ECAs the question arises of how referring expressions in which linguistic information and gestures are combined should be generated automatically, but also how such multimodal REs are produced by humans (Beun & Cremers, 1998; Byron, 2003).

The generation of referring expressions (GRE) is a central task in Natural Language Generation (NLG), and various algorithms which automatically produce REs have been developed (recent examples include Van Deemter & Krahmer, 2007; Van Deemter, 2002, 2006; Gatt 2006; Jordan & Walker, 2005; Gardent, 2002; Krahmer, et al. 2003). Existing GRE algorithms generally assume that both speaker and addressee have access to the same information. In most cases this information is represented by a knowledge base that contains the objects and their properties present in the domain of conversation. A typical algorithm takes as input a single object (**the target**) and a set of objects (**the distractors**) from which the target

object needs to be distinguished (borrowing terminology from Dale & Reiter, 1995). The task of a GRE algorithm is to determine which set of properties is needed to single out the target from the distractors. This is known as **content determination** for REs. On the basis of this set of properties a **distinguishing description** in natural language can be generated; a description which applies to the target but not to any of the distractors.

In general, there are multiple distinguishing descriptions for a given target. Consider, for instance, the chess configuration in Figure 1, with a circle around the target. This target can be described exclusively with linguistic features that express, say, that the target is a knight and that it is white. However, "the white knight" is not uniquely identifying, because there are two white knights on the board. Consequently, more or other information is needed to distinguish the target knight. For instance, the expression "the white knight" can be extended with several relational properties: "in row 5", "at position E5", "that is threatened by a black pawn", etc. Alternatively, a **multimodal referring expression** may be used, consisting of a pointing gesture plus a linguistic description such as "this knight". Arguably, such a multimodal RE would be easier to process than an overlong linguistic description, certainly when the addressee is a beginning chess player who is perhaps not familiar yet with the structure of the board or with the names of the various pieces.

<<Insert Figure 1>>

This paper presents a novel algorithm that generates **multimodal REs**. Pointing gestures are unique to human behavior and almost inevitable in human communication (c.f., Kita, 2004; Mc Neill, 1992). In contrast to the diversity of pointing gestures occurring in human conversation (e.g., Clark & Bangerter, 2004; Kendon, 2004; Kendon & Versante, 2003), the discussion in this paper is limited to pointing gestures that are performed by a hand with an

extended index finger that causes a projection of a straight line from the tip of the index finger to the intended referent. Various algorithms for the generation of multimodal REs have been proposed (c.f., Kranstedt & Wachsmuth, 2005; Lester et al. 1999; Andre & Rist, 1996; Claassen, 1992; Reithinger, 1992)[1] . Most of these are based on the assumption that a pointing gesture is precise and unambiguous. As soon as a pointing gesture is included, it directly eliminates the distractors and singles out the target. Consequently, the generated expressions tend to be relatively simple and usually contain no more than a head noun together with a pointing gesture. Moreover, most algorithms tend to be based on relatively simple, context-independent criteria for the decision whether a pointing gesture should be included or not.

In contrast, the GRE algorithm that is presented in this paper discards some of these basic assumptions. The multimodal GRE algorithm is based on observations in human communication and it approaches the generation of REs as a compositional task in which language and gestures are combined in a natural way. A key feature in the proposed model is the principle of Minimal Cooperative Effort (Clark & Wilkes-Gibbs, 1986) stating that both the speaker's effort in producing the description and the addressee's effort in interpreting it should be minimal. When considering this notion of effort, the multimodal GRE algorithm might decide to use a pointing gesture to indicate a target, as an undemanding way of identification. In a domain with many similar objects a purely linguistic description might be too complex to produce by a speaker and to interpret by an addressee.

Another assumption that is not retained by the algorithm presented in this paper is that a pointing gesture always has to be precise. The model for pointing that is proposed here allows for pointing gestures of different levels of precision, and likens a pointing gesture to the cone of a flashlight. If one holds a flashlight just above a surface (precise pointing), it covers only a

small area (the target object). Moving the flashlight away (imprecise pointing) enlarges the cone of light (shining on the target object but probably also on one or more other objects). A direct consequence of this Flashlight Model for pointing is that the amount of linguistic properties required to generate a distinguishing multimodal RE is predicted to co-vary with the kind of pointing gesture used. Consequently, the multimodal algorithm might decide to identify the target in Figure 1 with a precise pointing gesture combined with the linguistic description "this knight". Alternatively, the algorithm could use a less precise pointing gesture which also includes other objects in its scope, in which case more linguistic information is necessary to ensure a distinguishing description. For instance, assuming the scope of the pointing gesture to include the black knight at F6, the accompanying linguistic description at least should contain the property "white", as in "the white knight", to single out the target.

Precise pointing has a high precision. Its scope is restricted to the target object, and this directly rules out the distractors. But, arguably, precise pointing is "expensive"; the speaker has to make sure she points precisely to the target object in such a way that the addressee will be able to unambiguously interpret the RE. Imprecise pointing, on the other hand, has a lower precision ---it generally includes some distractors in its scope--- but is intuitively less "expensive". This intuition is in line with the alleged existence of neurological differences between precise and imprecise pointing. The former is argued to be monitored by a slow and conscious feedback control system, while the latter is governed by a faster and non-conscious control system located in the center and lower-back parts of the brain (e.g., Smyth & Wing, 1984; Bizzi & Mussa-Ivaldi, 1990).

In the computational model presented in this paper, the decision to point is based on a trade-off between the costs of pointing and the costs of a linguistic description. The latter are

determined by summing over the costs of the individual linguistic properties used in the description. Arguably, the costs of precise pointing are determined by two factors: the size of the target object (large objects are easier to point to than small objects) and the distance between the target object and the pointing device (objects that are near are easier to point to than objects that are further away). Below, Fitts' law (a fundamental empirical law about the human motor system attributable to Fitts, 1954) is used to model the costs of pointing. It is argued that Fitts' law allows the model to capture the intuition that imprecise pointing is cheaper than precise pointing.

In this paper it is shown how multimodal REs as appearing in cooperative human communication can be generated automatically. The next section presents a multimodal GRE algorithm as an extension of the graph-based algorithm proposed by Krahmer et al. (2003). The algorithm integrates the Flashlight Model for pointing in order to combine linguistic material and pointing gestures dependent on a notion of effort. Next we describe two production experiments that were conducted to evaluate this multimodal GRE algorithm. Finally, in the discussion, the results of this study are taken as a starting point to discuss ongoing work that aims to improve the algorithm's adaptation to human communication.

A Computational Model for the Generation of Multimodal Referring Expressions

Overview

The algorithm described in this paper is a multimodal variant of the graph-based GRE algorithm described by Krahmer et al. (2003). The algorithm represents the domain of conversation as a labeled directed graph, to which we refer as the **domain graph**. The objects in a domain are modeled as the vertices (or nodes) in the graph. To generate a RE for a target object, the graph-based algorithm searches for a subgraph of the domain graph that uniquely

represents the target. Which solution is returned depends on the cost function used. The **cost function** can be used to assign weights to the edges that represent the properties and relations, thereby determining their order of preference.

The graph-based approach has several advantages for GRE. One advantage is that there are many well-known search algorithms already in existence that deal with graph structures (e.g., Liebers, 2001; Messmer & Buncke, 1995, 1998; Eppstein, 1999). With the various existing search strategies in combination with proper cost functions, the graph-based algorithm can mimic existing GRE algorithms, such as those proposed by Dale (1989) and Dale & Reiter (1995). In this paper, we argue that the graph-based approach also lends itself well to the generation of multimodal REs. For that purpose, the domain graph is enriched with edges representing various kinds of pointing gestures. Since the algorithm presented here looks for the cheapest subgraph, pointing edges are selected only when linguistic edges are relatively expensive or when pointing is relatively cheap.

Generating Multimodal Referring Expressions Using Graphs

Domain Graphs

Consider the example domain depicted in Figure 2, consisting of a set of objects with various properties and relations.[2] For this particular domain, we have a the set of eight objects ($D = \{d_1,\ldots, d_8\}$), a set of properties of these objects (*Prop* = {*black*, *white*, *pawn*, *knight*, *bishop*, *rook*}) and a set of relations between these objects (*Rel* = {*left-of*, *right-of*}). This domain can be modeled as a labeled directed graph (which we will call the **domain graph**). The formal definition of domain graphs can be found in the appendix, where we will also list a number of other technical definitions. Our example domain can be represented as the graph in Figure 3. Not

all possible spatial relations are modeled in this domain graph under the assumption that a distinguishing description will not use a distant object as a relatum when a closer one can be selected.[3] Notice that properties are represented as loops, while relations are modeled as edges between different vertices.

<<Insert Figure 2>>

<<Insert Figure 3>>

Referring Graphs

Suppose that given the example domain, a distinguishing description referring to d_4 has to be generated. Then it has to be determined which properties and/or relations are required to single out d_4 from its distractors. This is done by creating **referring graphs,** which at least include a vertex representing the target object. Informally, a vertex *v* (the target object) in a referring graph refers to a given object in the domain graph if and only if the referring graph can be "placed over" the domain graph in such a way that *v* can be "placed over" the vertex of the given object in the domain graph and each edge from the referring graph labeled with some property or relation can be "placed over" a corresponding edge in the domain graph with the same label. Furthermore, a vertex-graph pair is distinguishing if and only if it refers to exactly one vertex in the domain graph. The informal notion of one graph being "placed over" another corresponds with a well-known mathematical construction on graphs, namely **subgraph isomorphism** (defined in the appendix).

Consider Figure 4 containing a number of potential referring graphs for d_4, where the vertex denoting d_4 is circled. The first one, H_1 has all the properties of d_4 and hence can refer to d_4. It is not distinguishing, however: it can also be placed over d_7 (the other large black knight), and thus fails to rule out this distractor. Graph H_2 is distinguishing. Here, the referring graph can

only be placed over the intended referent d_4 in the domain graph. A straightforward linguistic realization can be "a black knight to the left of a white pawn and to the right of a white pawn".[4] Generally there is more than one distinguishing graph referring to an object. In fact, H_2 is not a minimal distinguishing graph referring to d_4. This is H_3, which might be realized as "the black knight to the right of a pawn". This is a distinguishing description but not a particularly natural one; it is complex and arguably difficult for the addressee to interpret. In such cases, having the possibility of simply pointing to the intended referent would be very useful.

<<Insert Figure 4>>

Gesture Graphs

Suppose a pointing gesture is directed at d_4. Clearly this can be done from various distances and under various angles. The various hands in Figure 5 illustrate three levels of deictic pointing gestures, all under the same angle but each with different distances to the target object: **precise pointing** (***P***), **imprecise pointing** (***IP***) and **very imprecise pointing** (**VIP**). Here the presentation is limited to these three levels of precision and a fixed angle, although nothing hinges on this. Naturally, the respective positions of the speaker and the target object co-determine the angle under which the pointing gesture occurs; this in turn fixes the scope of the pointing gesture and thus which objects are ruled out by it (namely, those objects fully outside of the scope). If these respective positions are known, then computing the scope of a pointing gesture is straightforward; but the actual mathematics falls outside the scope of this paper (but see Kranstedt et al. 2006). Here we assume that, based on the positions of the target object and the speaker, the algorithm is able to compute the scope of a pointing gesture and a gesture graph is constructed accordingly.

<<Insert Figure 5>>

Just as properties and relations of objects can be expressed in a graph, so can various pointing gestures to these objects. All objects in the scope of a potential pointing gesture (with a certain degree of precision) are associated with an edge labeled with an indexed pointing gesture. Selecting this edge implies that all objects that fall outside the scope of the gesture are ruled out. This information is represented using a **gesture graph** (defined in the appendix). Figure 6 displays a graph modeling the various pointing gestures in Figure 5. Notice that there is one gesture edge which is only associated with d_4, the one representing precise pointing to the target object (modeled by edge P_{d4}). No other pointing gesture eliminates all distractors.

<<Insert Figure 6>>

Multimodal Graphs

Now the generation of multimodal referring graphs is based on the combination of the domain graph (which is relatively fixed) with the deictic gesture graph (which varies with the target). To generate a multimodal RE for a target object v, the graph-based algorithm first has to construct the gesture graph F_v associated with that target object, in order to produce the **multimodal graph** $M = F_v \cup G$ (formal definition in the appendix). Thus, M represents the search space of the multimodal GRE algorithm. As noted before, the search for a subgraph that uniquely describes the target object depends on the cost function used. A cost function assigns weights to the labeled edges in the graph. In the case of a multimodal graph both the costs of linguistic edges and the costs of gesture edges have to be determined. In the next section the cost functions for both kinds of edges are discussed.

Cost Functions

In the graph perspective there are many ways to generate a distinguishing RE for an object. Cost functions are used to give preference to some solutions over others. Costs are

associated with all subgraphs of the domain graph. The cost function is required to be **monotonic**. This implies that extending a graph with an edge can never result in a graph which is cheaper than the original graph. We shall assume that if *H* is a subgraph of *G*, the cost of *H* (notated Cost (*H*)) can be determined by summing over the costs associated with the vertices and edges of *H*.

The Costs of Properties and Relations

The cost of a subgraph is dependent on the costs associated with the edges in the graph. There are numerous ways in which costs can be assigned to edges. For instance, a cost function might simply associate each edge with a 1 point cost. In that case, when searching for the cheapest subgraph the algorithm will output the smallest distinguishing subgraph, which leads to the generation of minimal descriptions. Another approach is to define a cost function that models the notion of preferred attributes by Dale & Reiter (1995); for empirical evidence see e.g., Beun & Cremers (1998). In object descriptions people generally tend to include *type* properties. If that does not suffice, first absolute properties like *color* may be used, followed by relative ones such as *size*. A more fine-grained cost function might even differentiate between costs within one kind of property. For instance, *black* can be cheaper than *ebony*, if *black* is considered more common than *ebony* (cf. the basic level values as proposed by Dale & Reiter, 1995 and Krahmer & Theune, 2002). In terms of costs, we may assume that the *type* property is for free (costs no points), whereas other properties are more expensive, with absolute properties assumed to be cheaper than relative ones. There is little empirical work on the cost of relations, but it seems safe to assume that for the chess domain relations are more expensive than properties. Relations are comparable to relative properties in that they cannot be verified on the basis of the intended

referent alone. In addition, using a relation implies that a second object, the relatum, needs to be described as well and describing two objects generally requires more effort than describing a single object.

The Cost of Pointing

Arguably, at least two factors co-determine the cost of pointing: (1) The *size* of the target object (the larger the object, the easier, and hence cheaper, the pointing gesture); and (2) The *distance* which the pointing device (in this case the hand) has to travel in the direction of the target object (a short distance is cheaper than a long distance). Interestingly, the pioneering work of Fitts (1954) captures these two factors in the **Index of Difficulty** (*ID*), which states that the difficulty to reach a target is a function of the size, or the width *W*, of the target and the distance to the target, or amplitude *A*. With each doubling of *Distance* and with each halving of *Size* the Index of Difficulty increases by 1 bit. In recent years various alternatives for the original *ID* have been proposed. Here we use the alternative proposed by MacKenzie (1991), which starts counting from 1, ensuring that the *ID* is always positive:

$$ID = \mathrm{Log}_2 (A / W + 1)$$

As argued, it seems a reasonable assumption that imprecise pointing is cheaper than precise pointing; it rules out fewer distractors, but also requires less motor precision and effort from the speaker. In our domain, we can model this intuition using the Index of Difficulty in the following way. *Distance* is not interpreted as the distance from the current position of the hand to the target object, but rather as the distance from the current position of the hand to the target position of the hand. Thus, the smaller the distance from the current position of the hand to the target position for pointing, the lower the cost. In sum: if *g* is a pointing gesture, *A* is the distance from the

hand's current position to its target position, and *W* is the size of target object, then the cost associated with that pointing gesture is defined as follows:

$$\textbf{Costs}(g) = \text{Log}_2 (A / W + 1)$$

Worked Example

This section illustrates the algorithm with an example. The algorithm outputs the cheapest distinguishing graph for a target object, if one exists. Whether this cheapest graph will include pointing edges, and if so, of what level of precision, is determined by a trade-off between the respective costs of pointing and the costs of the linguistic edges. Interestingly, Piwek (2007) argues, based on a corpus study, that this trade-off may very per speaker (some speakers point easily, others only use pointing as a last resort). In terms of costs this means that for some speakers the relative costs of pointing are low (these speakers are thus likely to point), whereas for other speakers the relative costs for pointing are high.

Suppose our goal is to generate a description for d_4 from the scene graph *G* in Figure 3. To illustrate the workings of this function a specified cost function is needed. Let us assume that the distance from the current neutral position of the hand to the target position required for a precise pointing gesture to the target d_4 is 15cm, 7cm for imprecise pointing and 3cm for very imprecise pointing. When, for the sake of simplicity, the size of the target is assumed to be 1cm, some easy calculations will show that the index of difficulty in the three cases is 4 bits, 3 bits and 2 bits respectively. Thus, precise pointing (*P*) costs 4 points, imprecise pointing (*IP*) costs 3 points and very imprecise pointing (*VIP*) has cost 2 points.[5] Finally, let us assume for the sake of illustration that type edges (i.e. *knight*) are for free, color edges cost 1 point and relational edges 2 points.

Figure 7 sketches the multimodal generation algorithm; and here we will apply it to d_4 (i.e. we call the function **makeReferringExpression**(d_4, G)). First of all the deictic gesture graph F_{d4}, adding pointing edges of various levels of precision to d_4, is constructed, and merged with G. This results in the multimodal graph M. The variable *bestGraph*, for the cheapest solution found so far, is initialized as the undefined graph $\perp$ (no solution has been found yet), and the referring graph under construction H is initialized as the graph only consisting of the vertex d_4. Subsequently, the function **findGraph** is called with as parameters the target object (d_4), the best graph so far ($\perp$), the referring graph under construction (H) and the multimodal graph (M). Now the algorithm systematically tries all relevant subgraphs H of M. It starts from the graph which contains only the vertex d_4 and recursively tries to extend this graph by adding *adjacent* edges (i.e., edges which start in d_4 or possibly in any of the other vertices added later on to the H under construction). For each graph H it checks which objects in M (different from d_4) the vertex-graph pair (d_4, H) may refer to ("can be placed over"); these are the *distr*actors. As soon as this set is empty, a distinguishing graph referring to d_4 has been found. This graph is stored in the variable *bestGraph*, the cheapest distinguishing graph found so far. In the end the algorithm returns the cheapest distinguishing graph which refers to the target, if one exists, otherwise it returns the undefined null graph $\perp$. In the current setup the latter possibility will never arise due to the presence of unambiguous pointing gestures (expensive though they may be).

<<Insert Figure 7>>

Which referring graph is the first to be found depends on the order in which the edges are tried (clearly this is a place where heuristics are helpful, e.g. it will generally be beneficial to try cheap edges before expensive ones). Let us say, for the sake of argument, that the first distinguishing graph which the algorithm finds is H_2 from Figure 8. This graph costs 3 points. At

this point, graphs which are as expensive as this graph can be discarded (since due to the monotonicity constraint they will never end up being cheaper than the best solution found so far). In the current situation, the cheapest solution is H_1 from Figure 8, which costs 2 points. The resulting graph could be realized as "this knight" combined with a very imprecise pointing gesture. Note that if pointing were cheaper (because the distance between the current position of the hand and the required position for precise pointing was, say, 3cm), the algorithm would output "this knight" plus a precise pointing edge (i.e., H_3 from Figure 8, for cost 2 points). If pointing were more expensive (because even for very imprecise pointing the distance would be substantial), the algorithm would output "the black knight to the right of a white pawn" (i.e., H_2 from Figure 8, for cost 3 points).

<<Insert Figure 8>>

Experiment I

Overview

To what extent does the output of this algorithm resemble the expressions produced by human speakers? Evaluating content determination algorithms for natural language generation systems is known to be difficult. Corpora are often used for the evaluation of other natural language processing applications, but are not straightforwardly applicable to the evaluation of content determination algorithms, since typically the underlying semantic representations are not accessible (e.g. van Deemter et al., 2006). The descriptions extracted from corpora provide no information about the objects described, nor about their context. In this paper, production experiments are proposed for the evaluation of multimodal NLG algorithms. In such experiments, participants are offered various targets to which they have to refer. In this way, spontaneous data is gathered (i.e. participants are not told what to say), while controlling the

input representations at the same time (i.e. the target, its distractors and their respective properties are known). It can then be determined to what extent the verbalized output of the algorithm coincides with the utterances of the participants in the dimension under investigation. Such data is essential for data-driven development and testing of multimodal interpretation and generation modules (e.g. Kranstedt et al., 2006; Piwek & Beun, 2001).

Experiment I addresses one of the crucial ingredients of the algorithm: the claim that the linguistic part of a multimodal RE co-varies with the kind of pointing gesture. Given that the algorithm predicts that precise pointing rules out all distractors whereas an imprecise pointing gesture does not uniquely distinguish the target, this experiment was designed to test the hypothesis that when the distance to the target is small (i.e. pointing is precise) the linguistic description that accompanies the pointing gesture is relatively simple (i.e. at most includes a head noun). In contrast, when the distance to the target is relatively large (i.e. imprecise pointing) it is expected that participants combine their pointing gesture with an extensive linguistic expression to indicate the target. Although it seems likely that imprecise pointing requires more linguistic material to single out the target object, it is not known what kind of material is used. The experimental data will be used to provide more insight in this issue. Moreover, it might be that the complexity of the target object plays a role in the kind of linguistic descriptions that are combined with the various pointing gestures, where we expect that referring expressions for complex targets to contain more properties and relations than descriptions of simple targets.

Participants

Twenty native speakers of Dutch participated in the experiment, all students and colleagues from Tilburg University. None was familiar with the multimodal generation algorithm being tested. For each condition five men and five women participated.

Experimental setting

Participants had to perform an object identification task, in which they were first shown an isolated object which they subsequently had to single out among a set of comparable objects. Two sorts of target objects (geometrical figures and photos of persons) were used to determine whether the kind of target influenced the results. Half of the participants performed the tasks in the **near condition,** at a close distance (i.e. they could touch the target object directly and thus point precisely), the other half of the participants performed the same tasks in the **far condition**, from a distance of about 2.5m from the screen, from which they could only indicate the rough location of the target via imprecise pointing. Participants were led to believe they were testing a new computer system which could be operated by the combined usage of speech and gesture. They were told the system was in its testing phase; their input was required for calibration purposes. To evoke pointing gestures, the participants were given a pen-like **digital stick**, a pen mouse of approximately 10 centimeters, which could be used as a pointing device. They were told that the digital stick emitted a signal which the computer could detect and interpret. In addition, participants were equipped with a headset including a microphone through which they could speak to the computer. Their task was to identify the target objects via speech and gesture. Each target was first displayed in isolation on a 17inch screen, after which the target was presented among a set of distractors from which the participant had to single it out. To avoid influencing the participants in their realizations, no feedback was given by the experimenter or the computer.

Stimuli

Two kinds of target objects were used in the experiment: (1) 15 two-dimensional geometrical objects; and (2) 15 black and white photographs of persons (all famous

mathematicians). To facilitate pointing, the objects were presented on the screen in two isolated groups of 2 or 3 objects, one containing the target, the **target group**, while the other group solely consisted of distractors, the **distractor group**. The position of the target group on the screen was systematically varied, as was the position of the target object within the target group. Figures 9 and 10 illustrate the stimuli for objects and persons respectively. The geometrical figures vary in shape (square, circle, triangle) and color (red, blue, green). The persons display a greater variety: some are male, some female, they may wear hats, glasses, moustaches and/or beards (only the men), and they may have long, short, grey or no hair. A representative subset of 30 target objects was selected and presented to participants in a random order. For the identification task, the target object was presented on a computer screen together with a number of other objects from the same domain.

Data processing and statistical analysis

The participants were filmed during the experiment. The resulting data consist of (20 participants * 30 stimuli) = 600 multimodal REs. All utterances were transcribed. The kind of pointing gesture was classified, and the kinds of linguistic properties were determined and counted. All participants produced a correct, i.e. distinguishing, description for each target object. The descriptions were analyzed with respect to the following features:

- *Number of words* Per target description, the number of words used is counted.
- *Number of disfluencies* Per target description, the number of repairs, repetitions, pauses and filled pauses is counted.
- *Occurrences of type* In Dutch, *type* properties are mostly head nouns that describe the target. In a block domain these head nouns typically express the shape of an

object (e.g. “triangle”, “square”, “ball”). This feature counts the number of *type* properties used to describe the target.

- *Occurrences of properties* Per target description, the number of verbalized target properties (with the exception of *type*) are counted (e.g. “round”, “green”).
- *Number of locative relations* Per target description, the number of locative relations is counted (e.g. “on the left side”).

For each of the features an analysis of variance (ANOVA) with repeated measures was performed, with distance (levels: near, far) as between-subject variable and target (levels: objects, persons) as within-subject variable.

Results

We first checked whether the participants produced the intended pointing gestures. Indeed, all participants always used a pointing gesture. In the near condition, this pointing gesture was always a precise one, where the target object was directly touched with the pointing device. In the far condition, all participants produced imprecise pointing gestures, which basically denote in which of the two groups of objects on the screen the target object was located. This indicates that the operation of (im)precise pointing worked as planned, and the hypothesis can be tested that the kind of pointing gesture co-varies with the linguistic RE. No gender differences were found, so combined results for male and female participants are presented.

As a first approximation of speaker effort, the number of words is considered together with the number of disfluencies in the multimodal REs as a function of the distance and the target. The results are presented in Table 1. For both the number of words and the disfluencies there is a significant effect of distance (F (1,18) = 45.45, $p < .01$) and (F (1,18) = 9.24, $p < .01$),

respectively, which indicates that in the far condition participants use more words and less fluent speech than in the near condition. For the number of words there is also a significant effect of target (F (1,18) = 53.99, $p < .01$); this implies that participants require more words to refer to the persons than to the objects. In addition, there is an interaction between distance and target for both variables (words = F (1,18) = 49.09, $p < .01$ and disfluencies = F (1,18) = 3.48, $p < .08$). This can be explained by observing that the effect of distance is stronger for persons than for objects in the far condition but not in the near condition.

<<Insert Table 1>>

So, it appears that the linguistic material and the kind of pointing gesture used by speakers are co-related. Although some differences among the participants were observed, especially in the near condition, each of the participants displayed consistent behavior throughout the experiment. In the near condition, a precise pointing gesture suffices to single out the target object. Half of the participants in this condition only used a precise pointing gesture, three participants typically accompanied the gesture with a demonstrative determiner, “deze” (this one), the remaining two participants tended to include some more words in their multimodal REs. In the far condition, all participants used imprecise pointing gestures, and hence were required to use additional linguistic material to produce unambiguous REs.

Table 2 presents a more detailed analysis of the linguistic material, making a distinction between *type* information (whether the target is a square, a circle, person, etc., i.e., the information given in the head noun), the number of prenominal properties (*prop*) e.g., *color*, *hair style*, *hair color*, etc. and the number of locative relations (*loc*) e.g., *left*, *below*, etc. Looking at the presence of the property *type*, a significant effect of distance is found (F (1,18) = 144.6, $p < .01$); no effect of target and no interaction either (in both cases F (1,18) < 1). That is, when

participants use a precise pointing gesture in this experiment they do not use type information, but when they use an imprecise pointing gesture, they do include type information (sometimes even twice, explaining the 1.01 for persons). For adjectival properties, both a significant effect of distance is found (F (1,18) = 70.01, $p < .01$), and a significant effect of target (F (1,18) = 10.31, $p < .01$). No interaction is found. In terms of the figures in Table 2: when participants use a precise pointing gesture, they do not use adjectival properties, and when they use an imprecise pointing gesture they do. In addition, when participants describe an object they are somewhat more likely to use a prenominal adjective than when describing a person. For locations, finally, a significant effect of distance is found (F (1,18) = 2.02, $p < .05$), and a significant effect of target (F (1,18) = 20.47, $p < .01$). There is also an interaction between target and distance (F (1,18) = 16.62, $p < .01$). Inspection of the table reveals that these effects can be explained by the fact that location information is rare when a precise pointing gesture is used, but relatively common when describing a person in combination with an imprecise pointing gesture.

<<Insert Table 2>>

Summary

The experimental results indicate that speakers indeed vary the linguistic part of a multimodal RE in relation to the distance from the target object; the amount of linguistic material co-varies with the kind of pointing gesture. In the near condition, eight out of ten speakers always produced multimodal REs containing a demonstrative determiner, “deze” (this), or no spoken material at all. The remaining two consistently added a head noun, e.g. “deze driehoek” (this triangle). When, on the other hand, an imprecise pointing gesture is used, because of the distance to the target, the REs contain much more spoken material. The kind of target also had an influence on this. In general, fewer words are required to single out a geometrical figure than to

identify a person, in the current experiment. Closer inspection of the data reveals that both objects and persons are described in terms of their type (e.g. “triangle” and “man” respectively). In addition, geometrical objects are more often accompanied by prenominal adjectives (e.g., “blue”), while person descriptions tend to include locative relations (e.g. “at the top left”). This is probably due to the fact that describing persons is inherently more difficult than colored geometrical objects, since the number of potentially relevant attributes is much larger for persons than for geometrical objects.

One disadvantage of Experiment I is that participants were forced to point, so there is some risk that the produced multimodal referring expressions are not entirely representative of human multimodal references (but see Kühnlein & Stegmann 2003 for comparable results obtained in a study where the use of pointing gestures was not compulsory). In addition, the size of the target objects was kept constant, and participants were interacting with a computer instead of with another human being.

Experiment II

Overview

Experiment II has a similar set-up as Experiment I: participants again had to perform an identification task, but this time they were prompted by the experimenter to locate countries on a world map. Arguably, this is an inherently more spatial task but speakers were not forced to point. For this experiment, we made a distinction between targets that were ‘easy’ to point to (countries that are large or are isolated) and targets that were ‘difficult’ to point to (small and surrounded by distractors). We hypothesized that speakers would point more often to the ‘easy’ targets. Especially when referring to ‘difficult' targets, ‘easy’ relata could be helpful in identifying the target.

Participants

Twenty native speakers of Dutch participated in the second experiment (none was a participant in Experiment I). None was familiar with the multimodal generation algorithm. Ten participants were male and ten female, evenly distributed over the conditions.

Experimental setting

Participants were told that their topographical knowledge was going to be tested. Half of the participants performed the experiment in the **near condition** (they could directly touch the map), and half in the **far condition** (about 2.5 meters from the map). Participants in the far condition could only point imprecisely to a subarea of the map. All participants had to locate 30 countries, presented to them in a random order. Participants were given a stick of 40 cm which they could use for pointing if they so desired, but they were not explicitly instructed to point. Figure 11 shows two representative stills from the two experimental conditions.

Stimuli

30 countries were selected, which were assumed to be easy to locate on a world map for our participants. Of the selected countries, 15 were large (easily identifiable, e.g. Australia, Brazil, Canada) and 15 were small (difficult to identify, e.g. Belgium, Portugal, Surinam).

Data processing and statistical analyses

The participants were filmed during the experiment. The resulting corpus consist of (20 participants * 30 stimuli) = 600 multimodal REs. A typical example of a description of an easy object like Brazil is “dat grote groene vlak daar” (that large green area over there) together with an imprecise pointing gesture. A somewhat extreme example of reference to a difficult object (Portugal) is “het eh groene land dat ten zuid westen of dat eh in Zuid Europa ligt naast het roze Spanje” (the uh green country which lies in the south west or which uh lies in Southern Europe

next to the pinkish Spain) together with an imprecise pointing gesture. All utterances were orthographically transcribed. All participants produced a correct, i.e. distinguishing, description for each target object. The descriptions were analyzed using the same features as described for Study I (i.e., Number of words, number of disfluencies, occurrences of *type*, *properties* and *locative relations*). Two additional linguistic features were analyzed:

- *Occurrences of name* Per target description the times the name of the target is mentioned is counted. For instance “Brazil” in the example above.
- *Number of relata* Per target description, the number of relata used is counted. For instance, “Europa” and “Spanje” in the description of Portugal above. The descriptions that identify the relata are dealt with separately.

In addition the number of pointing gestures directed to the target and directed to relata were counted. For each of the features an analysis of variance (ANOVA) with repeated measures is performed, with distance (levels: near, far) as between-subject variable and target (levels: easy, difficult) as within-subject variable.

Results

We first checked when and where participants used pointing gestures, and it turned out that, without being explicitly instructed to do so, all participants always used a pointing gesture. In the near condition, this pointing gesture was always a precise one, where the target was directly touched. In the far condition participants could only use imprecise pointing gestures, which basically denote in what area on the map the target is located. This indicates that the variation in distance worked as intended. In Table 3 an analysis of the occurrences of is presented. The total number of pointing gestures (*total pointing*) is split into pointing to the target (*to target*) and pointing *to relata*. Table 3 shows that all participants pointed at every target at least once, no

matter the distance or size. When we consider the number of *total pointing gestures* in more detail, it appears that participants in the far condition direct considerably more pointing gestures *to relata* in describing difficult objects than in describing easy objects. More specifically, the total number of pointing gestures displays both an effect of distance (F (1,18) = 24.52, $p < .01$) and an effect of target (F (1,18) = 13.45, $p < .01$). Moreover the interaction between target and distance (F (1,18) = 11.62, $p < .01$) indicates that participants in the far condition use more pointing gestures in references to difficult objects.

<<Insert Table 3>>

Table 4 gives number of words and the number of disfluencies in the multimodal referring expressions for the various conditions. For the number of words there is an effect of distance (F (1,18) = 241.04, $p < .01$), and an effect of target (F (1,18) = 33.12, $p < .01$). These effects indicate that in the far condition participants use more words than in the near condition and participants require more words to refer to difficult objects than to easy ones. In addition, there is an interaction between distance and target (F (1,18) = 23.93, $p < .01$). This can be explained by observing that the effect of target is stronger in the far condition than in the near condition. The number of disfluencies show an effect of distance (F (1,18) = 100.44, $p < .01$) and an effect of target (F (1,18) = 6.44, $p < .05$), which indicate that participants speak less fluently in the far condition when referring to difficult objects. Furthermore there is an interaction between distance and target (F (1,18) = 7.17, $p < .05$) which signals a stronger effect of target in the far condition compared to the near condition. In the near condition participants do not use many words to refer to objects easy or difficult, consequently disfluencies are scarce.

<<Insert Table 4>>

Table 5 offers a more detailed analysis of the linguistic material, making a distinction between *name* (e.g., "Portugal"), *type* (e.g., head nouns like "country", "area", "part" etc.), the number of prenominal properties (e.g., *color*, *size*, *shape*, etc.) and the number of location markers (*location*). Location markers can be split into at least two types: (1) "in the south", as a general reference to the southern part of the world, and (2) "next to the pinkish Spain" including a relatum. In the latter case "next to the pinkish Spain" as a whole is treated as a location marker. In addition we counted the number of *relata* used per description. For a detailed analysis of the descriptions of the relata we refer to Van der Sluis (2005).

<<Insert Table 5>>

The results show that for almost all features there is a significant effect of distance (*name*, F (1,18) = 41.21, $p < .01$; *type*, F (1,18) = 132.21, $p < .01$; *property*, F (1,18) = 554.75, $p < .01$; *location*, F (1,18) = 76.57, $p < .01$; *relata*, F (1,18) = 119.787, $p < .01$). Thus, in the far condition, speakers use more names, more type, property and location information and more relata to identify a target object. The results also show that participants tend to use more *type* and *property* information when referring to easy targets (*type*, F (1,18) = 5.96, $p < .05$ and *property*, $F(1,18) = 5.94$, $p < .05$), whereas in descriptions for difficult targets participants use more *names*, *locations* and *relata* (*name*, F (1,18) = 5.03, $p < .05$; *location*, F (1,18) = 27.72, $p < .01$; *relata*, $F(1,18) = 51.157$, $p < .01$). When we compare the references for easy objects with those of difficult objects, it can be noted that the differences are almost non-existent in the near condition, while they are substantial in the far condition (*name*, F (1,18) = 4.91, $p < .05$; *type*, F (1,18) = 6.99, $p < .05$; *property*, F (1,18) = 9.53, $p < .05$; location, F (1,18) = 27.24, $p < .01$; *relata*, F (1,18) = 41.149, $p < .01$). Interestingly, in the far condition, easy objects are more often

referred to using head nouns and properties, while descriptions of difficult objects tend to contain more locative expressions and relata.

Summary

The results of this second experiment support the findings of the first one. Contrary to expectation, in the second study speakers always include pointing gestures in their descriptions regardless of the difficulty of the target and the distance to the target. This could be a result of the fact that the participants were equipped with a stick with which they could point, or simply because the nature of the task provokes pointing gestures. When the target is close, speakers tend to point only once in the direction of the target. Speakers use more pointing gestures to refer to small targets than to large targets. A closer inspection of the data shows that the extra pointing gestures are directed towards relata and not the target. Furthermore, speakers co-vary the linguistic part of a multimodal RE with the distance to the target and the kind of target. When the target is close, speakers reduce the linguistic material to almost zero (i.e. often indicating the target with a demonstrative determiner only), whereas participants tend to produce overspecified descriptions if the target is located further away. This can be due to the inherent uncertainty of imprecise pointing. Speakers may not be sure whether the imprecise pointing gesture is sufficiently clear and so, to guarantee that their reference is distinguishing, they include additional information. In human communication speakers typically produce REs that contain more information than strictly necessary to identify the target (e.g. Pechmann, 1989). Addressees appreciate this redundancy, because it facilitates comprehension of the speaker's message (cf. Engelhardt et al., 2006; Van der Sluis en Krahmer, 2005; Arts, 2004).

Discussion and Future work

This paper has described a new computational model for the generation of multimodal REs. The approach is based on only a few, independently motivated, assumptions. A Flashlight Model for pointing was proposed, allowing for different gradations of pointing precision, ranging from precise and unambiguous to imprecise and ambiguous. The algorithm used to generate the multimodal REs according to this model is a graph-based algorithm which tries to find the cheapest RE for a particular target object (Krahmer et al., 2003). In the search for the cheapest solution, it is assumed that linguistic properties have certain costs (cf. the preferred attributes from Dale & Reiter, 1995), whereas the costs of the various pointing gestures are derived from an empirically motivated adaptation of Fitts' law (Fitts, 1954). The costs should not be interpreted in an absolute sense, but relative to each other. In general, we assume that the relative weight of linguistic properties and deictic gestures is speaker-dependent (some speakers point more easily, and more frequently, than others).

The model has a number of nice consequences, such as (1) There is no a priori criterion needed to decide when to include a pointing gesture. Rather the decision to point is based on a trade-off between the costs of pointing and the costs of linguistic property; (2) The amount of linguistic properties required to generate a distinguishing multimodal RE is predicted to co-vary with the kind of pointing gesture; (3) An isolated object does not require precise pointing; there is always a graph containing a less precise and hence cheaper pointing edge which has the same objects in its scope as the more precise pointing gesture; and (4) A precise pointing edge and a relational edge never occur together in a distinguishing graph, because a graph that contains a precise pointing gesture is distinguishing.

To implement the Flashlight Model for pointing, the graph-based algorithm presents a very suitable framework. In contrast, an incremental strategy to the generation of multimodal descriptions is not straightforward. Such an incremental approach to generate multimodal REs was presented by Van der Sluis & Krahmer (2001) as a multimodal variant of the Incremental Algorithm (Dale & Reiter, 1995). There pointing gestures were generated only if a purely linguistic expression would be too complex, (i.e. the number of linguistic properties exceeds a certain threshold). In these cases, the generated linguistic RE was simply discarded and a precise pointing gesture was generated together with a simple RE which contained no more than a head noun. Naturally, this approach (first generating a linguistic description, and then possible discarding it for a gesture) is hardly incremental in the sense of Dale and Reiter. Another way of extending the Incremental Algorithm with the generation of pointing gestures would be to enrich the list of preferred attributes with the gestures *VIP*, *IP* and *P* (in that order of preference, modeling the increase in cost). In this approach, first a number of linguistic edges are selected followed by one or more pointing edges. However, the lack of backtracking of the incremental Algorithm entails that all selected properties would be realized, which implies that: (1) Multiple gestures might be generated; and (2) If *P* is generated it comes with more properties than necessary. This seems to suggest that the Flashlight Model is inherently non-incremental.

The graph-based algorithm presented in this paper does not use an a priori criterion to decide when to use a pointing gesture. The output modality is determined by a trade-off between the costs of pointing and the costs of a linguistic description, which have to be defined on an empirical basis.[6] Two assumptions underlie the algorithm: (1) the amount of linguistic information necessary to identify a target co-varies with the precision of the pointing gesture included; and (2) the linguistic information and pointing gesture depend on the kind and the size

of the target. Two production experiments were presented to evaluate the output of the graph-based algorithm. Experiment I was conducted in a very strict setting where the distance to the domain and the target objects were varied. Experiment II was conducted in a more natural setting, where both the distance to the objects and the size of the target objects were varied.

Overall, it may be concluded that the co-variation of the linguistic material and the precision of pointing gesture as predicted by the algorithm corresponds well with the results of the experiments. The multimodal algorithm agrees with the majority of speakers concerning the fact that the more precise the pointing gesture, the less linguistic material is generated to refer to an object. The resulting data also confirm the second assumption that underlies the algorithm; the linguistic part of the RE varies with the size and the kind of target. In cases where the target is difficult to describe because it is small or because it has a lot of features, speakers generally choose to use locative relations for identification. In contrast, when the target is easy to describe, because it is large or has features which are easy to distinguish, speakers use the intrinsic properties of the target for identification.

Although the algorithm performs fine when it comes to the above assumptions, there are some differences between the REs produced by human speakers and the ones generated by the algorithm. The most obvious one is that in the far (but not in the near) condition, humans tend to overspecify their references, while the algorithm always produces minimal REs. It is worth stressing though, that different search strategies are compatible with the graph-based perspective. Krahmer et al. (2003) illustrate this by describing a different search strategy which mimics the Incremental Algorithm and thus gives rise to a certain amount of redundancy. However, it can be argued that this kind of overspecification differs from human overspecification, in that it depends

on the interplay between the objects in the distractor set and the preference order of the attributes (see also Jordan, 2002 and Krahmer and Theune 2002).

One of the reasons why participants in the two experiments overspecify their references in the far condition may be that they are uncertain whether their REs clearly indicate the target and therefore choose to include potentially redundant information. Arguably, the addition of redundant properties and gestures increases the certainty of the speaker about the likelihood of a correct interpretation on the side of the addressee. As such the degree of overspecification seems tied to a notion of certainty; a speaker produces an overspecified RE to increase the probability that the addressee is able to interpret it correctly. An outline for such an approach in the context of multimodal references can be found in Van der Sluis (2005).

Acknowledgements

This work was done within the context of the TUNA project, funded by the EPSRC in the UK, under grant reference GR/S13330/01 and the IMOGEN project, funded by the Netherlands Organization for Scientific Research (NWO) in the IMIX programme. Special thanks for useful comments and suggestions are due to Harry Bunt, Kees van Deemter, Albert Gatt, Alfred Kranstedt, Fons Maes, Graeme Ritchie and Mariët Theune. Thanks to the three anonymous reviewers for their useful comments on the manuscript. Thanks also to the audiences at the various workshops and conferences that provided feedback on the preliminary versions of the work discussed in this paper: a preliminary version of the multimodal algorithm was presented at the ENLG'01, the evaluation studies were presented at the LREC'04, and the ICSLP'04 and preparatory work on overspecification was presented at the STD'05 workshop on dialogue modeling and generation.

References

Andre, E. & T. Rist (1996). Coping with temporal constraints in multimedia presentation Planning. In *Proceedings of the 13th Conference of the AAAI*, pp 142-147.

Arts, A. (2004). *Overspecification in Instructive Texts.* Ph. D. thesis, Tilburg University.

Beun, R. & A. Cremers (1998). Object reference in a shared domain of conversation. *Pragmatics & Cognition 6*(2), 121-152.

Bangerter, A. (2007). Pointing and describing in referential communication: When are pointing gestures used to communicate? In *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). Aberdeen, Scotland.*

Bizzi, E. & F. Mussa-Ivaldi (1990). Muscle properties and the control of the arm movement. In D. Osherson, S. Kosslyn & J. Hollerbach (Eds.), *Visual Cognition and Action*, Volume 2. MIT Press.

Byron, D. (2003). Understanding referring expressions in situated language, some challenges for real world agents. In *Proceedings of the 1st International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University.

Cassell, J., J. Sullivan, S. Prevost & E. Churchill (2000). *Embodied Conversational Agents.* MIT Press, Cambridge.

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

Clark, H. & Bangerter, A. (2004). Changing conceptions of reference. In I. Noveck & D. Sperber (Eds.), *Experimental Pragmatics* . Basingstoke, England: Palgrave Macmillan, pp. 25-49.

Claassen, W. (1992). Generating referring expressions in a multimodal environment. In R. Dale,

E. Hovy, D. Rösner & O. Stock (Eds.), *Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence, 587*, 247-262. Springer Verlag, Berlin.

Croitoru, M. & K. van Deemter (2007). A Conceptual graph approach to the generation of referring expressions. In *Proceedings of the 20th International Joint Conference on Artificial Intelligenc (IJCAI-2007)*, Hyderabad, India.

Dale, R. & E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science 18*, 233-263.

Van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics 28* (1), 37-52.

Van Deemter (2006). Generating referring expressions that involve gradable properties. To appear in *Computational Linguistics*, 2006.

Van Deemter, K. & E. Krahmer (2006). Graphs and booleans. In H. Bunt & R. Muskens (Eds.), *Computing Meaning*, Volume 3. Kluwer Academic Publishers.

Eppstein, D. (1999). Subgraph isomorphism in planar graphs and related problems. *Journal of Graph Algorithms and Applications 3* (3), 1-27

Engelhardt, P., K. Bailey & F. Ferreira (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 554-573.

Fitts, P. (1954). The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology 47*, 381-391.

Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the ACL*, Philidelphia, USA

Garey, W. & D. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-*

Completeness. W.H. Freeman & Company, New York.

Gatt, A. (2006). Structuring knowledge for reference generation: A clustering algorithm. *Proceedings of the 11th Meeting of the EACL*, Trento.

Horacek, H. (2005). Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th ENLG*, Aberdeen, UK

Jordan, P. (2002). Contextual influences on attribute selection for repeated descriptions. In K. van Deemter & R. Kibble (Eds.), *Information Sharing: References and Presupposition in Language Generation and Interpretation*, pp.295-328. CSLI Publications Stanford.

Jordan, P. & M. Walker (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research 24*, pp 157-194.

S. Kopp, B. Jung, N. Lessmann, I. Wachsmuth (2003). Max- A multimodal assistant in virtual reality construction. *KI-Kunstiche Intelligenz 4*, 11-17.

Lester, J., J. Voerman, S. Towns, & C. Callaway (1997). Deictic believability: Coordinating gesture, locomotion and speech in lifelike pedagogical agents. *Applied Artificial Intelligence 13* (4-5), 383-414.

Liebers, A. (2001). Planarizing graphs. *Journal of Graph Algorithms and Applications 5*(1), 1-74.

Kendon, A. (2004). *Gesture, Visible Actions as Utterance*. Cambridge University Press.

Kendon, A. & L. Versante (2003). Pointing by the hand in Neapolitan. In S. Kita (Ed.), *Pointing where Language, Culture and Cognition Meet.* pp. 109-137. Lawrence Erlbaum Associate Publishers, Manwah, New Jersey, London.

Kita, S. (2003). *Pointing where Language, Culture and Cognition Meet.* pp. 109-137. Lawrence Erlbaum Associate Publishers, Manwah, New Jersey, London.

Krahmer, E., S. van Erk & A. Verleg (2003). Graph-based generation of referring expressions. *Computational Linguistics 29* (1), 53-72.

Krahmer, E. & M. Theune (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pp. 223-246. CSLI Publications, Stanford.

Kranstedt, A. & I. Wachsmuth (2005). Incremental generation of multimodal deixis referring to objects. In *Proceedings of the 10th ENLG*, Aberdeen, UK.

Kranstedt, A., A. Lücking, T. Pfeiffer, H. Rieser, & I. Wachsmuth (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit & I. Wachsmuth (Eds.), *Situated Communication*, pp. 155-207. Mouton de Gruiter.

Kühnlein, P. & J. Stegmann (2003). *Empirical issues in deictic gesture: referring to objects in simple identification tasks*. Report 2003/3, SFB 360, University of Bielefeld

MacKenzie, I. (1991). *Fitts' Law as a Performance Model in Human-computer Interaction.* Ph.D. thesis, University of Toronto, Canada.

Mc Neill (1992). *Hand and Mind: What gestures reveal about thought*. University of Chicago Press, London, Chicago.

Messmer, B. & H. Buncke (1995). Subgraph isomorphism in polynomial time. Technical Report IAM 95-003, University of Bern, Institute of Computer Science and Applied Mathematics, Bern, Switzerland.

Oviatt, S (1999). Ten myths of multimodal interaction. *Communications of the ACM 42* (11), 74-81.

Paraboni, I & K. Van Deemter (2002). Generating Easy References: The case of document

deixis. In *Proceedings of INLG-2002*, New York, pp.113-119.

Paraboni, I, K. Van Deemter & J. Masthoff (2006). Overspecified reference in hierarchical domains: measuring the benefits for readers. In *Proceedings of the INLG-2006*, Sidney.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics 27*, 98-110.

Piwek, P. & R. Beun (2001). Multimodal referential acts in a dialogue game. From empirical investigations to algorithms. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 127-131.

Piwek, P. (2007). Modality choice for generation of referring acts: pointing versus describing. In *Proceedings of Workshop on Multimodal Output Generation (MOG 2007). Aberdeen, Scotland.*

Reithinger, N. (1992). The performance of an incremental generation component for multi-modal dialog contributions. In R. Dale, E. Hovy, D. Rösner & O. Stock (Eds.), *Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence, 587*, 263-276. Springer Verlag, Berlin.

Van Deemter, K., I. van der Sluis & A. Gatt (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Natural Language Generation Conference (INLG-2006)*, Sidney, Australia.

Van der Sluis, I. (2001). An empirically motivated algorithm for the generation of multimodal referring expressions. In *Proceedings of the 10th EACL meeting*, July 9-11, Toulouse France, p. 67-72.

Van der Sluis, I. & E. Krahmer. (2005). Towards the generation of overspecified multimodal

referring expressions. In *Proceedings of the Symposium on Dialogue Modelling and Generation at the 15th Annual Meeting of the Society for Text and Discourse (STD 2005)*, July 6-9, Amsterdam, the Netherlands

Van der Sluis, I. (2005). *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions*. Ph.D. thesis, Tilburg University, the Netherlands.

Smyth, M. & A. Wing (1984). *The Psychology of Human Movement*. Academic Press, New York.

Appendix: Graph Definitions

Labeled directed graph: Let $Labels = Prop \cup Rel$ be the set of labels with *Prop* and *Rel* disjoint, then $G = < V_G, E_G >$ is a labeled directed graph, where $V_G \subseteq D$ is the set of vertices and $E_G \subseteq V_G \times \text{Labels} \times V_G$ is the set of labeled directed edges.[7]

Subgraph isomorphism: $H = < V_H, \text{E}_\text{H}>$ can be **placed over** $G = <V_G, E_G >$ iff there exists a subgraph *G'* of *G* such that *H* is isomorphic to *G'*. *H* is **isomorphic** to *G'* iff there exists a bijection $\pi : V_H \rightarrow V_{G'}$, such that for all vertices $v, w \in V_H$ and all $l \in Labels$:

$$(v, l, w) \in \text{E}_\text{H} \leftrightarrow (\pi.v, l, \pi. w) \in E_{G'}$$

Given a graph *H* and a vertex *v* in *H*, and a graph *G* and a vertex *w* in *G*, it can be defined that the pair (*v*, *H*) **refers** to the pair (*w*, *G*) iff (1) *H* is a **connected graph**, i.e., each vertex has at least one edge that links it to another vertex; and (2) *H* is mapped to a subgraph of *G* by an isomorphism π and $\pi.v = w$. A vertex-graph (*v*, *H*) **uniquely refers** to (*w*, *G*) (i.e., (*v*, *H*) is **distinguishing**) iff (*v*, *H*) refers to (*w*, *G*) and there is no vertex *w'* in *G* different from *w* such that (*v*, *H*) refers to (*w'*, *G*).

Graph union: The **union** of graphs $H = < V_H, E_H >$ and $G = < V_G, E_G >$ is the graph $H \cup G = < V_H \cup V_G, E_H \cup E_G >$.

Graph extension: If $G = < V, E >$ is a graph and $e = (v, l, w)$ is an edge between vertices v and w and with label $l \in Labels$, then the **extension** of G with e (notated $G + e$) is the graph $< V \cup \{v, w\}, E \cup e >$.

Gesture graph: Let $Gest_v = \{P_v, IP_v, VIP_v\}$ be the set of deictic pointing gestures to a target object v. Then, given a domain graph $G = <V_G, E_G>$, a gesture graph $F_v = <V_G, E_F>$ is a labeled directed graph, where V_G is the set of vertices from the domain graph and $E_F = V_G \times Gest_v \times V_G$ the set of pointing edges. The subscript v in the gesture graph F_v indicates the target of the pointing gesture.

Multimodal graph: Let $Labels = Prop \cup Rel \cup Gest_v$ with $Prop$, Rel and $Gest_v$ disjoint. So $M = <V_M, E_M >$ is a labeled directed graph where $V_M \subseteq D$ is the set of vertices and $E_M \subseteq V_M \times Labels \times V_M$ is the set of labeled directed edges.

Computing costs: $\mathbf{Cost}(H) = \sum v \in V_H \; Cost(v) + \sum e \in E_H \; \mathbf{Cost}(e)$

Monotonicity of cost functions: $\forall H \subseteq G \; \forall_e \in E_G : \mathbf{Cost}(H) \leq \mathbf{Cost}\,(H + e)$

Footnotes

[1] These algorithms all operate on domains which are in the direct visual field of both speaker and addressee. Throughout this paper this assumption is made as well.

[2] For the sake of simplicity the examples used to illustrate this algorithm are restricted to a 2D domain with only a limited number of objects. This is not an inherent limitation of the algorithm.

[3] For a principled derivation of graphs from domain models, (see Croitoru & van Deemter 2007).

[4] A somewhat more involved realization module might realize this graph as “the black knight between the two white pawns”

[5] For the sake of simplicity the costs of the properties are chosen in such a way that easy calculations can be made, but nothing hinges on this.

[6] Standard techniques from data driven computational linguistics can be applied to find the best settings of the cost function. The usual strategy is to collect a large corpus and divide it in a training and a test part. The algorithm can be applied with a number of different settings to the training corpus and the best setting (e.g. the setting with the largest number of correct predictions) is then applied to the test set.

[7] Subscripts are omitted where this can be done without creating confusion.

Table 1

The average number of *words* and *disfluencies* per description and the *overall* means in Experiment I as a function of *distance* and *target*. Standard deviations between brackets.

			Distance	
			Near	*Far*
Target	*Object*	Words	.78 (1.21)	2.93(.87)
		Disfluencies	.00(.00)	.16(.35)
	Person	Words	.84(1.31)	5.45(1.32)
		Disfluencies	.01(.02)	.34(.25)
	Overall	Words	.81(1.22)	4.19(1.69)
		Disfluencies	.00(,02)	.25(.31)

Table 2

The average number of attributes *type, property* and *location* given per description and the *overall* means in Experiment I as a function of *distance* and *target.* Standard deviations between brackets.

			Distance	
			Near	*Far*
Target	*Object*	Type	.15(.32)	1.00(.00)
		Property	.19(.34)	.94(.13)
		Location	.09(.27)	.30(.43)
	Person	Type	.11(.17)	1.01(.04)
		Property	.03(.11)	.76(.26)
		Location	.12(.33)	.81(.45)
	Overall	Type	.12(.25)	1.00(0.3)
		Property	.11(.25)	.85(.22)
		Location	.11(.27)	.56(.50)

Table 3

The average number of pointing gestures given per description and the *overall* means in Experiment II as a function of *distance* and *target*. The *total* number of pointing gestures is divided in pointing gestures directed *to target* and *to relata*. Standard deviations between brackets.

			Distance	
			Near	*Far*
Target	*Easy*	Total	1.00(.00)	1.32(.22)
		To target	1.00(.00)	1.13(.14)
		To relata	.00(.00)	.20(.17)
	Difficult	Total	1.02(.04)	1.87(.59)
		To target	1.02(.05)	1.03(.21)
		To relata	.00(.00)	.85(.64)
	Overall	Total	1.03(.15)	1.06 (.12)
		To Target	1.01(.29)	1.08(.29)
		To Relata	.00(.00)	.05(.84)

Table 4

The average number of *words* and *disfluencies* per description and the *overall* means in Experiment II as a function of *distance* and *target*. Standard deviations between brackets.

			Distance	
			Near	*Far*
Target	*Easy*	Words	2.28 (1.09)	15.59(3.10)
		Disfluencies	.19(.10)	1.57(.85)
	Difficult	Words	3.23(1.63)	27.25(6.28)
		Disfluencies	.17(.11)	2.40(.65)
	Overall	Words	2.76(1.44)	21.42(7.68)
		Disfluencies	.18(.11)	1.98(.85)

Table 5

The average number of attributes *name, type, property, location* and *relata* given per description and the *overall* means in Experiment II as a function of *distance* and *target*. Standard deviations between brackets.

			Distance	
			Near	*Far*
Target	*Easy*	Name	.32(.26)	.84(.24)
		Type	.03(.05)	.92(.29)
		Property	.04(.06)	1.51(.20)
		Location	.12(.13)	.18(.68)
		Relata	.05(.08)	1.11(.46)
	Difficult	Name	.33(.28)	1.07(.18)
		Type	.04(.07)	.76(.16)
		Property	.07(.08)	1.30(.21)
		Location	.13(.18)	2.87(.98)
		Relata	.11(.15)	2.20(.56)
	Overall	Name	.32 (.27)	.96 (.24)
		Type	.04(.06)	.84(.24)
		Property	.05(.67)	1.40(.23)
		Location	.12(.15)	2.32(1.00)
		Relata	.08(.12)	1.65(.75)

Figure 1

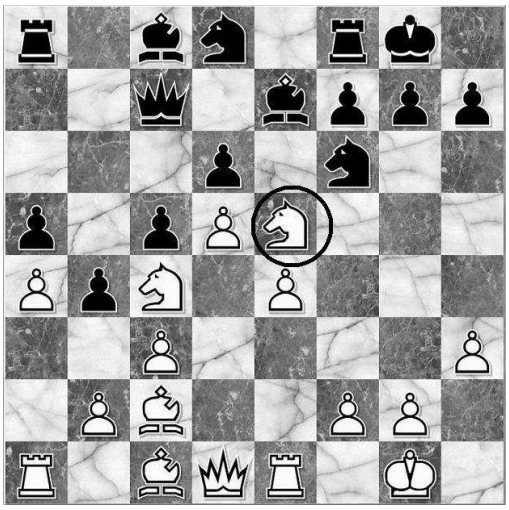

Figure 2

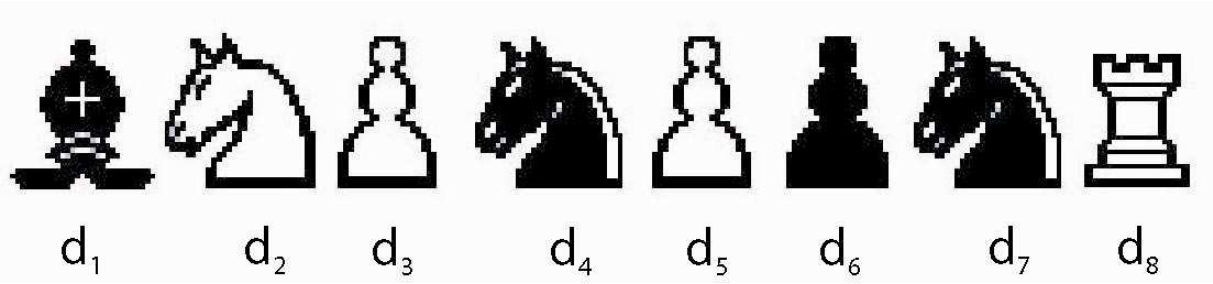

Figure 3

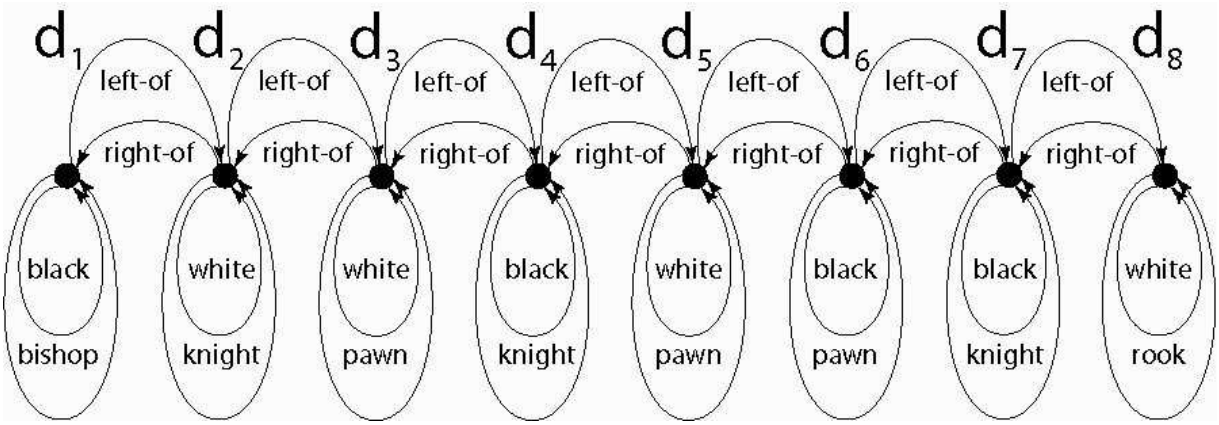
d1
d2
d3
d4
d5
d6
d7
d8
left-of
right-of
black
white
bishop
knight
pawn
rook

Figure 4

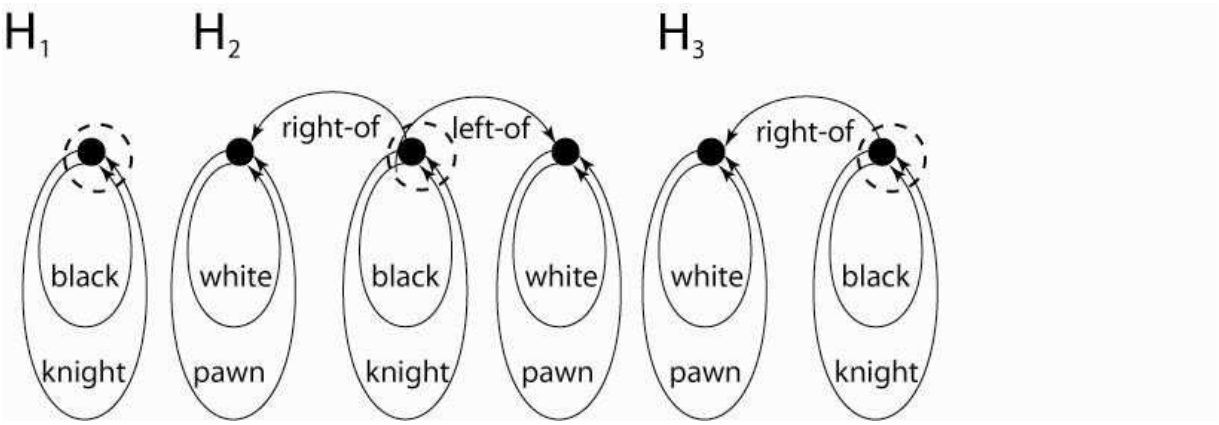
H1
H2
H3
right-of
left-of
right-of
black
knight
white
pawn
black
knight
white
pawn
white
pawn
black
knight

Figure 5

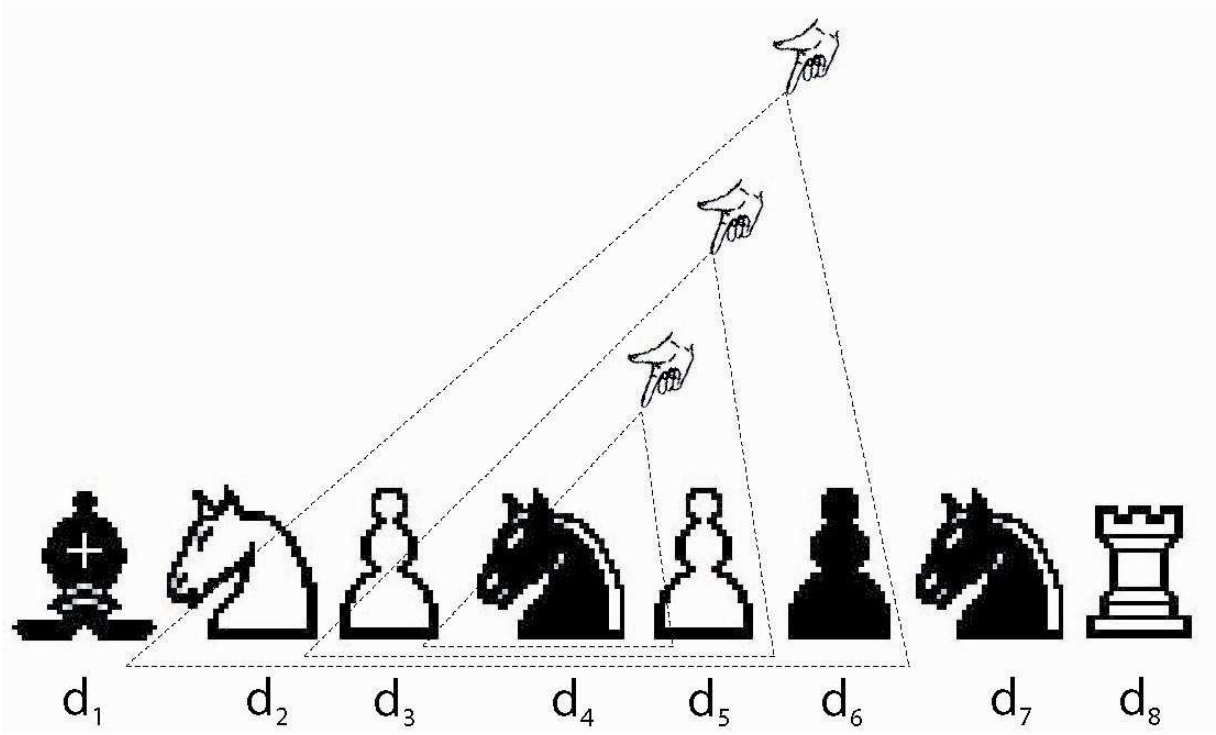

Figure 6

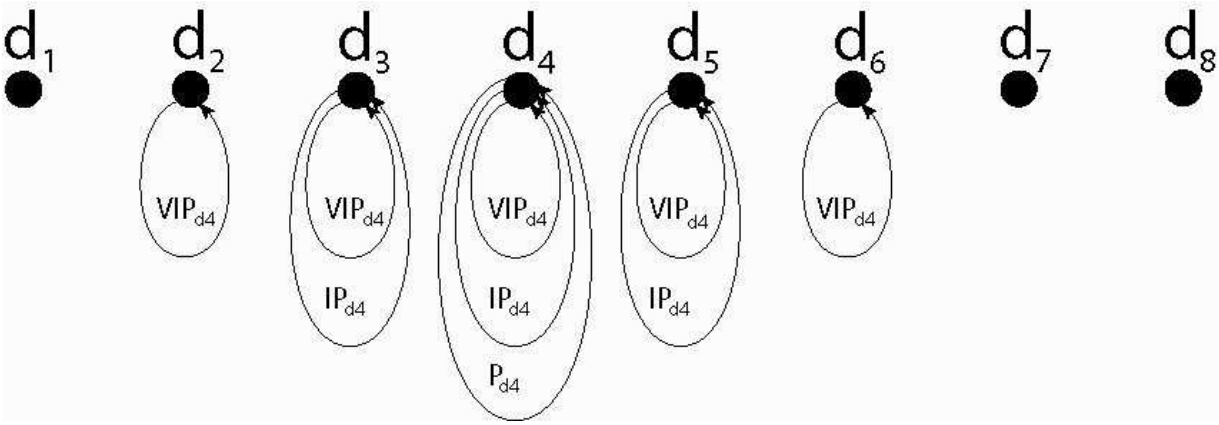
d1
d2
d3
d4
d5
d6
d7
d8
VIPd4
VIPd4
VIPd4
VIPd4
VIPd4
IPd4
IPd4
IPd4
Pd4

Figure 7

GenerateReferringExpression(*v, G*)
    Construct(*v*, F_v, *G*)
    $M := F_v \cup G$
    BestGraph : = $\perp$
    $H := \langle \{v\}, \phi \rangle$
    BestGraph : = **FindGraph**(*v*, *BestGraph*, *H*, *M*)
    Return *BestGraph*
FindGraph(*v*, *BestGraph*, *H*, *M)*
    if *BestGraph* $\neq \perp$ **and Cost**(*BestGraph*) $\leq$ **Cost**(*H)* **then**
        return *BestGraph*
    end if
    C : = $\{n \mid n \in V_M \wedge$ **MatchGraphs**(*v, H, n, M*)$\}$
    if *C* = $\{v\}$ **then**
        Return *H*
    end if
    for each adjacent edge *e* **do**
        I : = **FindGraph**(*v*, *BestGraph*, *H* + *e*, *M)*
        if *BestGraph* = $\perp$ **or Cost**(*I*) $\leq$ **Cost**(*BestGraph*) **then**
            BestGraph : = I
        end if
    end for each
    return *BestGraph*

Figure 8

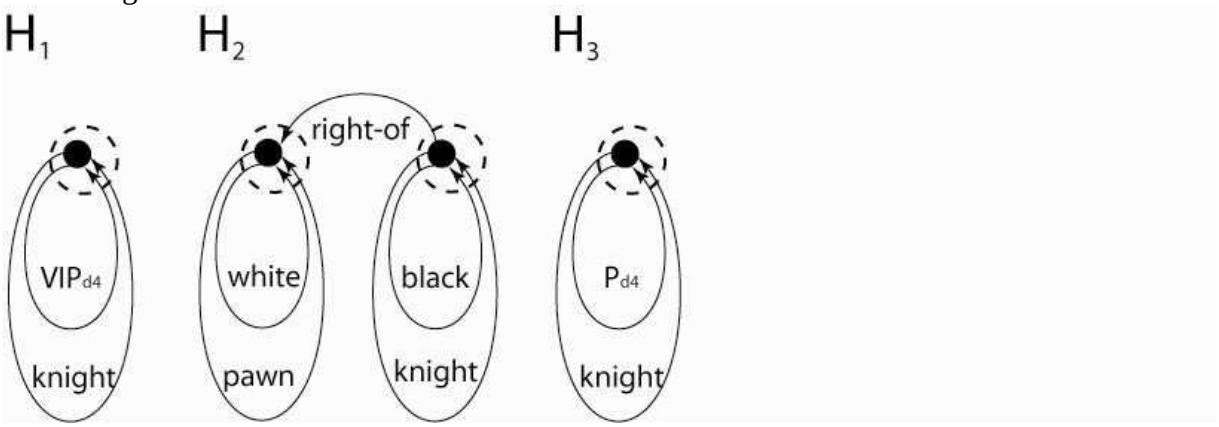
H1
H2
H3
VIPd4
knight
right-of
white
pawn
black
knight
Pd4
knight

Figure 9

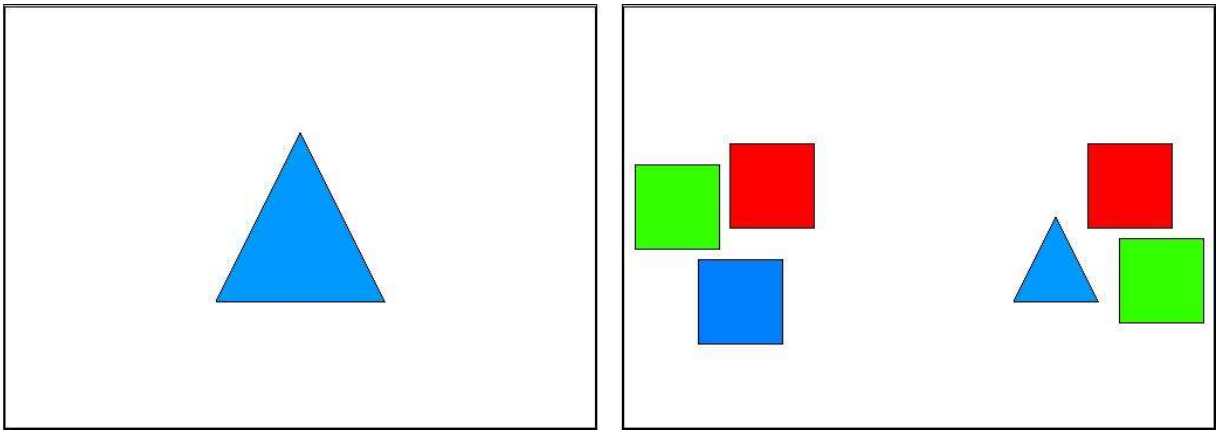

Figure 10

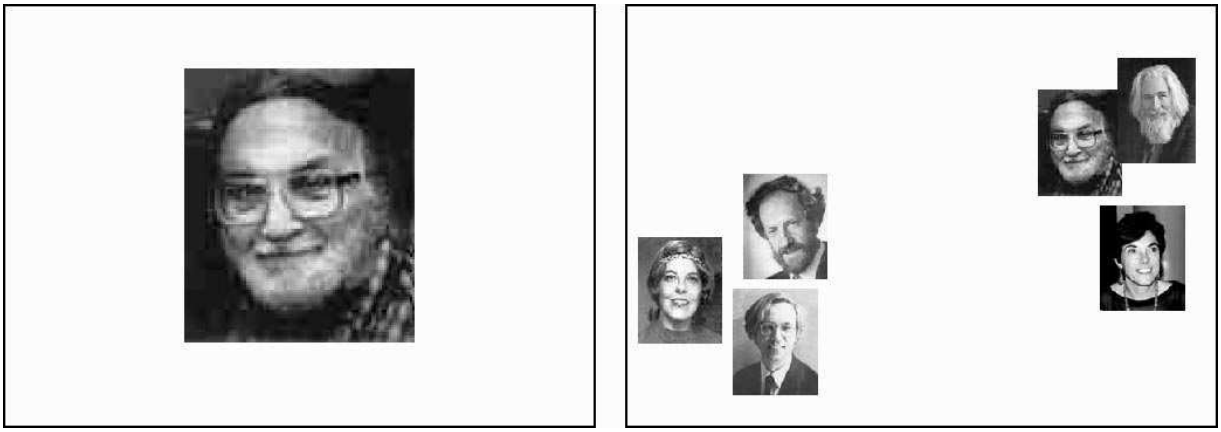

Figure 11

Figure Captions

Figure 1. A chess configuration.

Figure 2. An example domain.

Figure 3. Our example domain represented as a domain graph.

Figure 4. Three potential referring graphs for d_4 in the example domain.

Figure 5. Pointing into the example domain.

Figure 6. Deictic gesture graph.

Figure 7. Pseudocode of the algorithm's main function *generateReferringExpression* and the subgraph construction function *findGraph*.

Figure 8. Referring graphs for d_4 in the example domain.

Figure 9. A stimulus example from the domain of geometrical object. On the left, the target object displayed in isolation. On the right the target is presented with a number of similar objects.

Figure 10. A stimulus example from the domain of photographed persons. On the left, the target object (a picture of a mathematician) is displayed in isolation. On the right the target is presented together with a number of similar objects.

Figure 11. Example of participants in Experiment II in the near (left) and in the far (right) condition.