Natural Reference to Objects in a Visual Domain

Margaret Mitchell
Computing Science Dept.
University of Aberdeen
Scotland, U.K.

Kees van Deemter
Computing Science Dept.
University of Aberdeen
Scotland, U.K.

Ehud Reiter
Computing Science Dept.
University of Aberdeen
Scotland, U.K.

{m.mitchell, k.vdeemter, e.reiter}@abdn.ac.uk

Abstract

This paper discusses the basic structures necessary for the generation of reference to objects in a visual scene. We construct a study designed to elicit naturalistic referring expressions to relatively complex objects, and find aspects of reference that have not been accounted for in work on Referring Expression Generation (REG). This includes reference to object parts, size comparisons without crisp measurements, and the use of analogies. By drawing on research in cognitive science and psycholinguistics, we begin developing the input structure and background knowledge necessary for an algorithm capable of generating the kinds of reference we observe.

1 Introduction

One of the dominating tasks in Natural Language Generation (NLG) is the generation of expressions to pick out a referent. In recent years there has been increased interest in generating referential expressions that are *natural*, e.g., like those produced by people. Although research on the generation of referring expressions has examined different aspects of how people generate reference, there has been surprisingly little research on how people refer to objects in a real-world setting. This paper addresses this issue, and we begin formulating the requirements for an REG algorithm that refers to visible three-dimensional objects in the real world.

Reference to objects in a visual domain provides a straightforward extension of the sorts of reference REG research already tends to consider. Toy examples outline reference to objects, people, and animals that are perceptually available before the speaker begins generating an utterance (Dale and Reiter, 1995; Krahmer et al., 2003; van Deemter et al., 2006; Areces et al., 2008). Example referents may be referred to by their color, size, type (“dog” or “cup”), whether or not they have a beard, etc.

In theory, the target referent in an REG algorithm has a clearly defined set a properties, and the reference process proceeds by comparing these properties with the properties of other items in the set. The final expression roughly conforms to the Gricean maxims (Grice, 1975).

However, when the goal is to generate natural reference, this framework is too simple. The form reference takes is profoundly affected by modality, task, and audience (Chapanis et al., 1977; Cohen, 1984; Clark and Wilkes-Gibbs, 1986), and even when these aspects are controlled, different people will refer differently to the same object (Mitchell, 2008). In light of this, we isolate one kind of natural reference and begin building the algorithmic framework necessary to generate natural utterances within this domain.

Psycholinguistic research has examined reference in a variety of settings, which may inform research on natural REG, but it is not always clear how to extend this work to a computational model. This is true in part because these studies favor an analysis of reference in the context of collaboration; reference is embedded within language, and language is often a joint activity. However, most research on referring expression generation assumes a solitary generating agent, leaving no room for collaboration. This tacitly assumes that reference will be taking place in a monologue setting, rather than a dialogue or group setting. Indeed, the goal of most REG algorithms is to produce uniquely distinguishing, one-shot referring expressions.

Studies on natural reference usually use a two-person (speaker-listener) communication task (e.g., Flavell et al., 1968; Krauss and Glucksberg, 1969; Ford and Olson, 1975). This research has shown that reference is more accurate and efficient when it incorporates things like gesture and gaze (Clark and Krych, 2004). There is a trade-off in effort between initiating a noun phrase and refashioning it so that both speakers understand the referent (Clark and Wilkes-Gibbs, 1986), and speakers communicate to form lexical pacts on how to refer to an object (Sacks and Schegloff, 1979; Brennan and Clark, 1996). Mutual understanding of referents is achieved in part by referring within a subset of potential referents (Clark et al., 1983; Beun and Cremers, 1998). A few studies have compared monologue to dialogue reference, and have shown that monologue references tend to be harder for a later listener to disambiguate (Clark and Krych, 2004) and that subsequent references tend to be longer than those in dialogues (Krauss and Weinheimer, 1967).

Aiming to generate natural reference in a monologue setting raises questions about what structures an algorithm should use to produce utterances like those produced by people. In a monologue setting, the speaker (or algorithm) gets no feedback from the listener; the speaker's reference is not tied to interactions with other participants. The speaker is therefore in a difficult position, attempting to clearly convey a referent without being able to check if the reference is understood along the way.

Recent studies that have focused on monologue reference do so rather explicitly, which may affect participant responses. These studies utilize 2D graphical depictions of simple 3D objects (van Deemter et al., 2006; Viethen and Dale, 2008), where a small set of properties can be used to distinguish one item from another. The expressions are elicited in isolation, typed and then submitted, which may hide some of the underlying referential processes. None of these studies utilize actual objects. It is therefore difficult to use these data to draw conclusions about how reference works in less controlled settings. It is unclear if these experimental settings are natural enough, i.e., if they get at reference as it may occur everyday.

The study in this paper attempts to bring out information about reference in a number of ways. First, we conduct the study in-person, using real-world objects. This design invites referential phenomena that may not have been previously observed in simpler domains. Second, the referring expressions are produced orally. This allows us access to reference as it is generated, without the participants revising and so potentially obscuring information about their reference. Third, we use a relatively complicated task, where participants must explain how to use pieces to put together a picture of a face. The fact that we are looking at reference is not made explicit, which lessens any experimental effects caused by subjects guessing the purpose of the study. This approach also situates reference within a larger task, which may draw out aspects of reference not usually seen in experiments that elicit reference in isolation. Fourth, the objects used display a variety of different features: texture, material, color, size along several dimensions, etc. This brings the data set closer to objects that people interact with every day. A monologue setting offers a picture of the phenomena at play during a single individual's referring expression generation.

The referring expressions gathered in this study exhibit several aspects of reference that have not yet been addressed in REG. This includes (1) part-whole modularity; (2) size comparisons across three dimensions; and (3) analogies. Work in psycholinguistics, cognitive modelling, and neurophysiology suggests that these phenomena are interrelated, and may be possible to represent in a computational framework. This research also offers connections to further aspects of natural reference that were not directly observed in the study, but will need to be accounted for in future work on naturalistic referring expression generation. Using these ideas, we begin formulating the structures that an REG algorithm would need in order to produce reference to real-world objects in a visual setting.

Approaching REG in this way allows us to tie our research in the generation of referring expressions to work in computational models of visual perception and cognitively-motivated computer vision. Moving in this direction offers the prospect of eventually developing an application for the generation of natural reference to objects automatically recognized by a computer vision system.

In the next section, we describe our study. In Section 3, we analyze the results and discuss what they tell us about natural reference. In Section 4, we draw on our results and cognitive models of object recognition to begin building the framework for a referring expression algorithm that generates

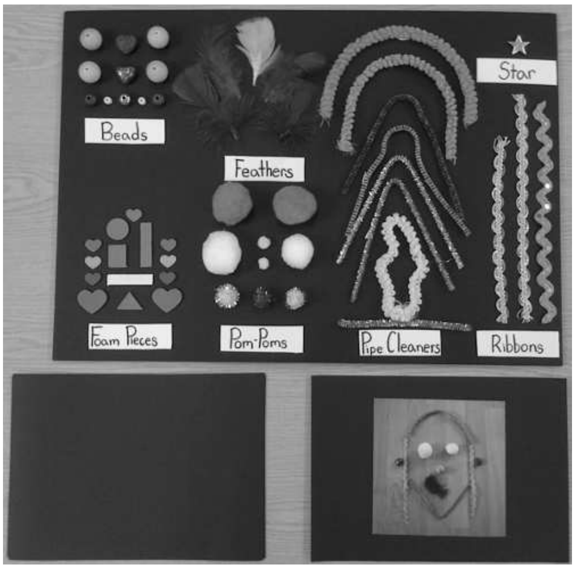

Figure 1: Object Board.

naturalistic reference to objects in a visual scene. In Section 5, we offer concluding remarks and outline areas for further study.

2 Method

2.1 Subjects

The subjects were 20 residents of Aberdeen, Scotland, and included undergraduates, graduates, and professionals. All were native speakers of English, had normal or corrected vision, and had no other known visual issues (such as color-blindness). Subjects were paid for their participation. Two recordings were left out of the analysis: one participant's session was not fully recorded due to a software error, and one participant did not pick out many objects in each face and so was not included. The final set of participants included 18 people, 10 female and 8 male.

2.2 Materials

A board was prepared with 51 craft objects. The objects were chosen from various craft sets, and included pom-poms, pipe-cleaners, beads, and feathers (see Table 1). The motley group of objects had different colors, textures, shapes, patterns, and were made of different materials. Similar objects were grouped together on the board, with a label placed underneath. This was done to control the head noun used in each reference. The objects were used to make up 5 different craft "face" pictures. Subjects sat at a desk facing the board and the stack of pictures. A picture of the board is shown in Figure 1.

Subjects were recorded on a head-mounted microphone, which fed directly into a laptop placed on the left of the desk. The open-source audio-recording program Audacity (Mazzoni, 2010) was used to record the audio signal and export it to wave format.

2.3 Procedure

Subjects were told to give instructions on how to construct each face using the craft supplies on the board. They were instructed to be clear enough for a listener to be able to reconstruct each face without the pictures, with only the board items in front of them. A pilot study revealed that such open-ended instructions left some subjects spending an inordinate amount of time on the exact placement of each piece, and so in the current study subjects were told that each face should take "a couple" minutes, and that the instructions should be as clear as possible for a listener to use the same objects in reconstructing the pictures without being "overly concerned" with the details of exactly how each piece is angled in relation to the other.

Subjects were first given a practice face to describe. This face was the same face for all subjects. They were then allowed to voice any concerns or ask questions, but the experimenter only repeated portions of the original instructions; no new information was given. The subject could then proceed to the next four faces, which were in a random order for each subject. A transcript of a single face from a session is provided in Figure 2.

2.4 Analysis

Each of the four face pictures used in the study was labelled with numbers for each craft item. All subject recordings were transcribed, including disfluencies, and each face section ("eyes", "chin", etc.) was marked. First reference to items on the board were annotated with their corresponding item numbers, yielding 722 references.[1] Initial references to single objects were extracted, yielding a final data set including 505 references to single objects.

[1] This corpus is available at http://www.csd.abdn.ac.uk/˜mitchema/craft_corpus.

14 foam shapes	2 large red hearts	2 small red hearts	2 small neon green hearts
2 small blue hearts	1 small green heart	1 green triangle	1 red circle
1 red square	1 red rectangle	1 white rectangle	
11 beads	4 large round wooden beads	2 small white plastic beads	2 brown patterned beads
1 gold patterned bead	1 shiny gold patterned heart	1 red patterned heart	
9 pom poms	2 big green pom-poms	2 small neon green pom-poms	2 small silver pom-poms
1 small metallic green pom-pom		1 large white pom-pom	1 medium white pom-pom
8 pipe cleaners	1 gold pipe-cleaner	1 gold pipe-cleaner in half	1 silver pipe-cleaner
1 circular neon yellow soft pipe-cleaner		1 neon orange puffy pipe-cleaner	1 grey puffy pipe-cleaner
1 purple/yellow striped pipe-cleaner		1 brown/grey striped pipe-cleaner	
5 feathers	2 purple feathers	2 red feathers	1 yellow feather
3 ribbons	1 gold sequined wavy ribbon	1 silver wavy ribbon	1 small silver wavy ribbon
1 star	1 gold star		

Table 1: Board items.

<CHIN> Okay so this face again um this face has um uh for the chin, it uses (10 *a gold pipe-cleaner in a V shape*) where the bottom of the V is the chin. </CHIN>
<MOUTH> The mouth is made up of (9 *a purple feather*). And the mouth is slightly squint, um as if the the person is smiling or even smirking. So this this smile is almost off to one side. </MOUTH>
<NOSE> The nose is uh (5 *a wooden bead, a medium-sized wooden bead with a hole in the center*). </NOSE>
<EYES> And the eyes are made of (2,3 *white pom-poms*), em just uh em evenly spaced in the center of the face. </EYES>
<FOREHEAD> Em it's see the person's em top of the person's head is made out of (1 *another, thicker pipe-cleaner that's uh a grey color, it's kind of uh a knotted blue-type pipe-cleaner*). So that that acts as the top of the person's head. </FOREHEAD>
<HAIR> And down the side of the person's face, there are (7,8 *two ribbons*) on each side. (7,8 *And those are silver ribbons*). Um and they just hang down the side of the face and they join up the the grey pipe-cleaner and the top um of the person's head to the to the chin and then hang down either side of the chin. </HAIR>
<EARS> And the person's ears are made up of (4,6 *two beads, which are um love-heart-shaped beads*), where the points of the love-hearts are facing outwards. And those are just placed um around same em same em horizontal line as the nose of the person's face is. </EARS>

Figure 2: Excerpt Transcript

3 Results

Each reference was annotated in terms of the properties used to pick out the referent. For example, "the red feather" was annotated as containing the <ATTRIBUTE:value> pairs <COLOR:red, TYPE:feather>. Discerning properties from the modifiers used in reference is generally straightforward, and all of the references produced may be partially deconstructed using such properties.

Using sets of properties to distinguish referents is nothing new in REG. Algorithms for the generation of referring expressions commonly use this as a starting point, proposing that properties are organized in some linear order (Dale and Reiter, 1995) or weighted order (Krahmer et al., 2003) as input. However, we find evidence that more is at play.

3.1 Spatial Reference

In addition to properties that pick out referents, throughout the data we see reference to objects as they exist in space. Size is compared across different dimensions of different objects, and reference is made to different parts of the objects, picking out pieces within the whole. These two phenomena – relative size comparisons and part-whole modularity – point to an underlying structural object representation that may be utilized during reference.

3.1.1 Relative Size Comparisons

A total of 122 (24.2%) references mention size with a vague modifier (e.g., "big", "wide"). This includes comparative (e.g, "larger") and superlative (e.g., "largest") size modifiers, which occur 40 (7.9%) times in the data set. Examples are given below.

(1) *"the bigger pom-pom"*

(2) *"the green largest pom-pom"*

(3) *"the smallest long ribbon"*

(4) *"the large orange pipe-cleaner"*

Of the references that mention size, 37 (7.3%) use a modifier that applies to one or two dimensions. This includes modifiers for height ("the short silver ribbon"), width ("quite a fat rectangle"), and depth ("the thick grey pipe-cleaner"). 87 (17.2%) use a modifier that applies to the overall size of the object (e.g., "big" or "small"). Crisp

Part-whole modularity	**Relative size**	**Analogies**
“a green pom-pom… with the tinsel on the outside” “your gold twisty ribbon… with sequins on it” “a wooden bead… with a hole in the center” “one of the green pom-poms… with the sort of strands coming out from it.” “the silver ribbon… with the chainmail detail down through the middle of it.”	“a red foam-piece… which is more square in shape rather than the longer rectangle” “the grey pipe-cleaner… which is the thicker one… “the slightly larger one” “the smaller silver ribbon” “the short silver ribbon” “quite a fat rectangle” “thick grey pipe-cleaner”	“a natural-looking piece of pipe-cleaner, it looks a bit like a rope” “a pipe-cleaner that looks a bit like… a fluffy caterpillar” “the silver ribbon that’s almost like a big S shape.” “a… pipe-cleaner that looks like tinsel.”
11 References	122 References	17 References

Table 2: Examples of Observed Reference.

measurements (such as “1 centimeter”) occur only twice (0.4%), with both produced by the same participant.

Participants produce such modifiers without sizes or measurements explicitly given; with an input of a visual object presentation, the output includes size modifiers. Such data suggests that natural reference in a visual domain utilizes comparison processes of the length, width, and height of a target object with other objects in the set. Indeed, 5 references (1.0%) in our data set include explicit comparison with the size of other objects.

(5) *“a red foam-piece… which is more square in shape rather than the longer rectangle”*

(6) *“the grey pipe-cleaner… which is the thicker one… of the selection”*

(7) *“the shorter of the two silver ribbons”*

(8) *“the longer one of the ribbons”*

(9) *“the longer of the two silver ribbons”*

In Example (5), height and width across two different objects are compared, distinguishing a square from a rectangle. In (6) “thicker” marks the referent as having a larger circumference than other items of the same type. (7) (8) and (9) compare the height of the target referent to the height of similar items.

The use of size modifiers in a domain without specified measurements suggests that when people refer to an object in a visual domain, they are sensitive to its size and structure within a dimensional, real-world space. Without access to crisp measurements, people compare relative size across different objects, and this is reflected in the expressions they generate. These comparisons are not only limited to overall size, but include size in each dimension. This suggests that objects’ size within a real-world space is relevant to REG in a visual domain.

3.1.2 Part-Whole Modularity

The role a spatial object understanding has within reference is further detailed by utterances that pick out the target object by mentioning an object part. 11 utterances in our data include mention of an object part within reference to the whole object. This is spread across participants, such that half of the participants make reference to an object part at least once.

(10) *“a green pom-pom, which is with the tinsel on the outside”*

(11) *“your gold twisty ribbon...with sequins on it”*

(12) *“a wooden bead...with a hole in the center”*

In (10), pieces of tinsel are isolated from the whole object and specified as being on the outside. In (11), smaller pieces that lay on top of the ribbon are picked out. And in (12), a hole within the bead is specified.

The use of part-whole modularity suggests an understanding that parts of the object take up their own space within the object. An object is not only viewed as a whole during reference, but parts in, on, and around it may be considered as well. For an REG algorithm to generate these kinds of reference, it must be provided with a spatial, structural representation for each object.

3.2 Analogies

The data from this study also provide information on what can be expected from a knowledge base in an algorithm that aims to generate naturalistic reference. Reference is made several times to objects not on the board, where the intended referent is compared against something it is *like*. For example, one participant makes reference to a SIZE property of an object not on the board: “two green pom-poms. . . they’re about the size of a ping-pong ball”. Other participants refer to objects that may share a variety of properties: “a natural-looking piece of pipe-cleaner, it looks a bit like a rope” and “a pipe-cleaner that looks a bit like. . . a fluffy caterpillar. . . ”

Reference to these other items do not pick out single objects, but types of objects (e.g., an object *type*, not *token*). They correspond to some prototypical idea of an object with properties similar to those of the referent. Work by Rosch (1975) has examined this tendency, introducing the idea of *prototype theory*, which proposes that there may be some central, ‘prototypical’ notions of items. A knowledge base with stored prototypes could be utilized by an REG algorithm to compare the target referent to item prototypes. Such representations would help guide the generation of reference to items not in the scene, but similar to the target referent.

Examples from the data of each of the observed phenomena are given in Table 2. Each column also lists the number of expressions with the given property.

4 Discussion

We have discussed several different aspects of reference in a study where referring expressions are elicited to objects in a spatial, visual scene. This is undoubtedly not an exhaustive account of the phenomena at play in such a domain, but offers some initial conclusions that may be drawn from exploratory work of this kind.

Before continuing with the discussion, we must first note that the argument can be made that our data does not exhibit reference – at least as it tends to be known in REG – but *description*. As discussed in the introduction, NLG systems favor a view where referring expressions pick out a target item in light of contextual alternatives. The idea that the target item may be picked out in some other way, for example, by the use of analogy or the inclusion of properties that do not rule out any alternatives, suggests that our subjects are not referring in the traditional REG sense. Perhaps they are doing something else, which could be called describing.

How to draw the line between a distinguishing reference and a description, and whether such a line can be drawn at all, is an interesting question. If the two are clearly distinct, then both are interesting to NLG research. If the two are one in the same, then this sheds some light on how REG algorithms should treat reference. We leave a more detailed discussion of this for future work, but note recent psycholinguistic work suggesting that referring establishes (i) an individual as the referent; (ii) a conceptualization or perspective on that individual (Clark and Bangerter, 2004). Schematically,

referring = indicating + describing.

We now turn to a discussion of how the observed phenomena may be best represented in an REG algorithm. Far from being a process that simply distinguishes a referent with regard to the other items in the set, natural reference to objects in a visual scene displays a variety of referential phenomena, focusing on object forms as they exist in a three-dimensional space and drawing on background knowledge to describe referents by analogy to items outside of the scene. We propose that an algorithm capable of generating natural reference to objects in a visual scene should therefore utilize (1) a spatial object representation; (2) a non-spatial feature-based representation; and (3) a knowledge base of object prototypes.

4.1 Spatial and Visual Properties

It is perhaps unsurprising to find reference that exhibits spatial knowledge in a study where objects are presented in three-dimensional space. Human behavior is anchored in space, and spatial information is essential for our ability to navigate the world we live in. However, referring expression generation algorithms geared towards spatial representations have oversimplified this tendency, keeping objects within the realm of two-dimensions and only looking at the spatial relations between objects.

For example, Funakoshi et al. (2004) and Gatt (2006) focus on how objects should be clustered together to form groups. This utilizes some of the spatial information between objects, but does not address the spatial, three-dimensional nature of objects themselves. Rather, objects exist as entities that may be grouped with other entities in a set or singled out as individual objects; they do not have their own spatial characteristics. Similarly, one of the strengths of the Graph-Based Algorithm (Krahmer et al., 2003) is its ability to produce reference to the spatial relations between objects (“next to”, “on top of”, etc.), but objects are essentially one-dimensional, represented as individual nodes.

Work that does look at the spatial information of individual objects is provided by Kelleher et al. (2005). In this approach, the overall volume of each object is calculated to assign salience rankings, which then allow the Incremental Algorithm to produce otherwise “underspecified” reference. The spatial properties of the objects are kept relatively simple. They are not used in constructing the referring expression, but one aspect of the object’s three-dimensional shape (volume) affects the referring expression’s surface form. To the authors’ knowledge, the current work is the first to suggest that objects themselves should have their spatial properties represented during reference.

Research in cognitive modelling supports the idea that we attend to the spatial properties of objects when we view them (Blaser et al., 2000), and that we have purely spatial attentional mechanisms operating alongside non-spatial, feature-based attentional mechanisms (Treue and Trujillo, 1999). These feature-based attentional mechanisms pick out properties commonly utilized in REG, such as texture, orientation, and color. They also pick out edges and corners, contrast, and brightness. Spatial attentional mechanisms provide information about where the non-spatial features are located in relation to one another, size, and the spatial interrelations between component parts.

Applying these findings to our study, a spatial representation in an REG algorithm that generates natural reference should utilize a dimensional representation of referent objects. This would provide information about the relative distances between object components, edges, and corners. Such a representation would enable the generation of size modifiers (“big”, “small”) without the need for crisp measurements, for example, by comparing the difference in overall height of the target object with other objects in the scene, or against a stored prototype. This would also allow for relative size comparisons across different dimensions, which may be used to generate size modifiers that refer to one dimensional axis, for example, “wide” and “thick”.

How we view and refer to objects appears to be influenced by the interaction between such a spatial representation and a property-based representation. Expectations about an object’s spatial properties guide our attention towards expected object parts and non-spatial, feature-based properties throughout the scene (Kosslyn, 1994; Itti and Koch, 2001). This affects the kinds of things we are most likely to generate language about (Itti and Arbib, 2005).

4.2 Analogies

Expectations about objects’ visual and spatial characteristics are derived from stored representations of object ‘prototypes’ in the inferior temporal lobe of the brain (Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 2000; Palmeri and Gauthier, 2004). Most formal theories of object perception posit some sort of *category activation system* (Kosslyn, 1994), a system that matches input properties of objects to those of stored prototypes, which then helps guide visual attention to expected object parts in a top-down fashion.[2] Such a representation stores objects’ component parts and where they are placed relative to one another, as well as relevant values for texture, material, color, etc. Features of a visual scene become salient (in part) if they deviate from the expected norms. This allows us to say, for example “the suitcase with the missing handle”, where expectations about what a suitcase looks like and where its handle should be guides the reference we produce.

This also allows us to say “the pipe-cleaner that looks like a caterpillar”, where knowledge about the shape, color, and form of a caterpillar influences the language we generate about objects that look similar to it. A prototype system appears to be a neurological correlate of the knowledge base we propose to underlie analogies. Matching properties of referents to items in the knowledge base allows object prototypes with a large number of shared properties to become relevant to reference.

A prototype system offers extensions to reference we have not directly observed in this study, but may want to account for later. For example, the expression “the three-legged dog” is likely to be produced to refer to a dog with three legs, whether or not there are any other items in the scene; such an expression may be generated when a compari-

[2] Note that this is not the only proposed matching structure in the brain – an *exemplar activation system* matches input to stored exemplars.

- A spatial representation (depicting size, interrelations between component parts)
- A non-spatial, propositional representation (describing color, texture, orientation, etc.)
- A knowledge base with stored prototypical object propositional and spatial representations

Table 3: Requirements for an REG algorithm that generates natural reference to visual objects.

son of the target referent to a dog prototype makes the number of legs salient.

We are now in a position to begin outlining some general requirements for an algorithm capable of generating naturalistic reference to objects in a visual scene: Input to such an algorithm should include a non-spatial feature-based representation, which we will call a *propositional representation*, with values for color, texture, etc., and a *spatial representation*, with symbolic information about objects' size and the spatial relationships between components. A system that generates naturalistic reference must also use a knowledge base storing information about object prototypes with their own propositional/spatial representations.

5 Conclusions and Future Work

We have explored the interaction between viewing objects in a three-dimensional, spatial domain and referring expression generation. This has led us to propose structures that may be used to connect vision in a spatial modality to naturalistic reference. These include a spatial representation, a propositional representation, and a knowledge base with spatial and propositional representations for object prototypes. Using structures that define the propositional and spatial content of objects in a scene fits well with work in psycholinguistics, cognitive modelling, and neurophysiology, and may provide the basis to generate a variety of natural-sounding references from a system that recognizes objects.

It is important to note that any naturalistic experimental design limits the kinds of conclusions that can be drawn about reference. For example, a study that elicits reference to objects in a visual scene provides insight into reference to objects in a visual scene; these conclusions cannot easily be extended to reference to other kinds of phenomena, such as reference to people in a novel. We therefore make no claims about reference as a whole in this paper, but reference within the particular domain we are testing. These generalizations can provide hypotheses for further testing in different modalities and with different sorts of referents.

Our data leave open many areas for further study, and we hope to address these in future work. Experiments designed specifically to elicit relative size modifiers, reference to object components, and reference to objects that are *like* other things would help further detail the form our proposed structures take.

What is clear from our data is that both a spatial understanding and a (non-spatial) property-based understanding appear to play a role in reference to objects in a visual scene, and further, reference in such a setting is bolstered by a knowledge base with stored prototypical object representations. Utilizing structures representative of these phenomena, we may be able to extend object recognition research into object reference research, generating natural-sounding reference in everyday settings.

Acknowledgements

Thanks to Advaith Siddarthan for thought-provoking discussions and to the anonymous reviewers for useful suggestions.

References

Carlos Areces, Alexander Koller, and Kristina Striegnitz. 2008. Referring expressions as formulas of description logic. *Proceedings of the Fifth International Natural Language Generation Conference*, pages 42–29.

Robbert-Jan Beun and Anita H. M. Cremers. 1998. Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6:121–52.

Erik Blaser, Zenon W. Pylyshyn, and Alex O. Holcombe. 2000. Tracking an object through feature space. *Nature*, 408:196–199.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–93.

Alphonse Chapanis, Robert N. Parrish, Robert B. Ochsman, and Gerald D. Weeks. 1977. Studies in interactive communication: II. the effects of four communication modes on the linguistic performance of teams during cooperative problem solving. *Human Factors*, 19:101–125.

Herbert H. Clark and Adrian Bangerter. 2004. Changing ideas about reference. In Ira A. Noveck and Dan Sperber, editors, *Experimental pragmatics*, pages 25–49. Palgrave Macmillan, Basingstoke, England.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Herbert H. Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22:1–39.

Philip R. Cohen. 1984. The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10(2):97–146.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

J. H. Flavell, P. T. Botkin, D. L. Fry Jr., J. W. Wright, and P. E. Jarvice. 1968. *The Development of Role-Taking and Communication Skills in Children*. John Wiley, New York.

William Ford and David Olson. 1975. The elaboration of the noun phrase in children's description of objects. *The Journal of Experimental Child Psychology*, 19:371–382.

Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. 2004. Generating referring expressions using perceptual groups. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 51–60.

Albert Gatt. 2006. Structuring knowledge for reference generation: A clustering algorithm. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 321–328.

Paul H. Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Laurent Itti and Michael A. Arbib. 2005. Attention and the minimal subscene. In Michael A. Arbib, editor, *Action to Language via the Mirror Neuron System*. Cambridge University Press.

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience*.

J. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102.

Stephen M. Kosslyn. 1994. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Robert M. Krauss and Sam Glucksberg. 1969. The development of communication: Competence as a function of age. *Child Development*, 40:255–266.

Robert M. Krauss and Sidney Weinheimer. 1967. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6:359–363.

Nikos K. Logothetis and David L. Sheinberg. 1996. Visual object recognition. *Annual Review Neuroscience*, 19:577–621.

Dominic Mazzoni. 2010. Audacity.

Margaret Mitchell. 2008. Towards the generation of natural reference. Master's thesis, University of Washington.

Thomas J. Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience*, 5:291–303.

Maximilian Riesenhuber and Tomaso Poggio. 2000. Models of object recognition. *Nature Neuroscience Supplement*, 3:1199–1204.

Eleanor Rosch. 1975. Cognitive representation of semantic categories. *Journal of Experimental Psychology*, 104:192–233.

Harvey Sacks and Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. In George Psathas, editor, *Everyday Language: Studies in Ethnomethodology*, pages 15–21. Irvington Publishers, New York.

Stegan Treue and Julio C. Martinez Trujillo. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–579.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation*, Sydney, Australia. ACL.

Jette Viethen and Robert Dale. 2008. The use of spatial descriptions in referring expressions. In *Proceedings of the 5th International Conference on Natural Language Generation, INLG-08*, Salt Fork, Ohio. ACL.