Attribute-Centric Referring Expression Generation

Robert Dale and Jette Viethen

Centre for Language Technology
Macquarie University
Sydney, Australia
rdale@science.mq.edu.au|jviethen@science.mq.edu.au

Abstract. In this chapter, we take the view that much of the existing work on the generation of referring expressions has focused on aspects of the problem that appear to be somewhat artificial when we look more closely at human-produced referring expressions. In particular, we argue that an over-emphasis on the extent to which each property in a description performs a discriminatory function has blinded us to alternative approaches to referring expression generation that might be better-placed to provide an explanation of the variety we find in human-produced referring expressions. On the basis of an analysis of a collection of such data, we propose an alternative view of the process of referring expression generation which we believe is more intuitively plausible, is a better match for the observed data, and opens the door to more sophisticated algorithms that are freed of the constraints adopted in the literature so far.

1 Introduction

Ever since at least the late 1980s, the generation of referring expressions has been a key focus of interest in the natural language generation community (see, for example, [1–12]). A glance at the proceedings of workshops in natural language generation over the last ten years demonstrates that the topic attracts significantly more attention than other aspects of the generation process, such as text structuring, sentence planning, and linguistic realisation; and, largely because of this critical mass of interest, the generation of referring expressions has served as the focus of the first major evaluation efforts in natural language generation (see, for example, [13, 14]).

This level of attention is due in large part to the consensus view that has arisen as to what is involved in referring expression generation: the task is widely agreed to be one that involves a process of selecting those attributes of an intended referent that distinguish it from other potential distractors in a given context, resulting in what is often referred to as a *distinguishing description*. Based on this agreement on the nature of the task, a large body of work has developed over the last 20 years that has focused on developing algorithms that encompass an ever-wider range of referential phenomena. Key issues that have

underpinned much of this work are the need to take account of the computational complexity of the algorithms developed; the production of descriptions which are in some sense minimal (in that they do not contain unnecessary information); and, occasionally, a recognition of some of the phenomena that characterise the kinds of referring expressions that humans produce.

Our key point is that the last of these concerns has not occupied the central position that it should, and that the other criteria that have been considered in order to determine what counts as a good algorithm have been given undue weight. Our position is based on two observations. First, it is clear (and this observation is not new) that humans do not always produce what are referred to as *minimal distinguishing descriptions*, i.e. referring expressions whose content walks the line between being both necessary and sufficient, despite this having served as a concern for much algorithmic development in the past. As has long been recognised, human-produced referring expressions are in many cases informationally redundant. The Incremental Algorithm [3], which serves as the basis for many algorithmic developments in the literature, is occasionally given credit because it can lead to referring expressions that contain redundancy; but even its authors were careful not to claim that the redundancy it produces is the same as that produced by humans. The kinds of redundancy evident in human-produced referring expressions have never, in our view, been properly explored, and this has led to algorithms which at best pay lip-service to the need to account for redundancy.

Our second observation (also not particularly new, but surprisingly ignored in the literature) is that different people do different things when faced with the same reference task. This poses serious questions for both the development of algorithms and their evaluation: as has been noted for other tasks that involve natural language output (such as document summarisation), in such circumstances we clearly cannot evaluate an algorithm by comparing its results against a single gold-standard answer. Even with a range of possible candidate answers, it is still possible that an algorithm might produce a perfectly acceptable solution that is not present amongst this set. This forces us to consider more carefully what it is that we are doing when we develop algorithms for the generation of referring expressions (or, for that matter, for any generation task): are we trying to emulate or predict the behavior of a single given speaker in a given situation? Or are we trying to produce a solution which might somehow rate as optimal in a task-based evaluation scenario (such as might be measured by the amount of time it takes a listener to locate a referred-to object), recognizing that human-produced referring expressions are not necessarily optimal in this sense? What counts as a good solution may be quite different in each case.

In this chapter, we examine some human-produced data in order to observe the variety that it exhibits (Section 2). We then posit a different way of thinking about the process of referring expression generation (Section 3), and go on to demonstrate how some machine-learning experiments run on the human-produced data are supportive of this view (Section 4). Although this way of looking at the problem is, we argue, more explanatory than previous approaches,

it still leaves a number of important questions unanswered; we discuss these in Section 5, before drawing some conclusions in Section 6.

2 What Do People Do?

For the purposes of the explorations discussed in this chapter, we use a corpus of human-produced referring expression data called the GRE3D3 Corpus. This corpus is introduced and discussed in significant detail elsewhere (see [15, 16]); here we summarise its key characteristics.

The data in question was gathered via a web-based experiment where participants were asked to produce referring expressions that would enable a listener to identify one of a number of objects shown on the screen. The purpose of the experiment was to explore how relations were used in referring expressions, and the design of scenes was carefully controlled so that the use of relations was encouraged but not strictly necessary in order to identify the intended referent.

Participants visited a website, where they first saw an introductory page with a set of instructions and a sample stimulus scene. The task was to describe the target referent in the scene (marked by a grey arrow) in a way that would enable a friend looking at the same scene to pick it out from the other objects. Figure 1 shows an example stimulus.

Each participant was assigned one of two trial sets of ten scenes each; the two trial sets are superficially different (involving colour variations and mirror-image orientations), but the elements of the sets are pairwise identical in terms of the factors explored in the research. The complete set of 20 scenes is shown in Figure 2: Trial Set 1 consists of Scenes 1 through 10, and Trial Set 2 consists of Scenes 11 through 20.[1] Each scene contains three objects, which we refer to as the *target* (the intended referent), the potential *landmark* (a nearby object), and the *distractor* (a further-away object).

The experiment was completed by 74 participants from a variety of different backgrounds and ages. One participant asked for their data to be discarded. We also disregarded the data of one other participant who reported to be colour-blind. One participant consistently produced very long and syntactically complex referring expressions including reference to parts of objects and the onlooker, such as *the red cube which rests on the ground and is between you and the yellow cube of equal size*. While these descriptions are very interesting, they are clearly outliers in our data set.

Eight participants consistently only used `type` to describe the target object, for example simply typing *cube* for the target in Scene 5. These descriptions were excluded from the corpus under the assumption that the participants had not understood the instructions correctly or were not willing to spend the time

[1] Scene 1 is paired with Scene 11, Scene 2 with Scene 12, and so on; in each pair, the only differences are (a) the colour scheme used and (b) the left–right orientation, with these variations being introduced to make the experiment less monotonous for participants; in [15], we report that these variations appear to have no significant effect on the forms of reference used.

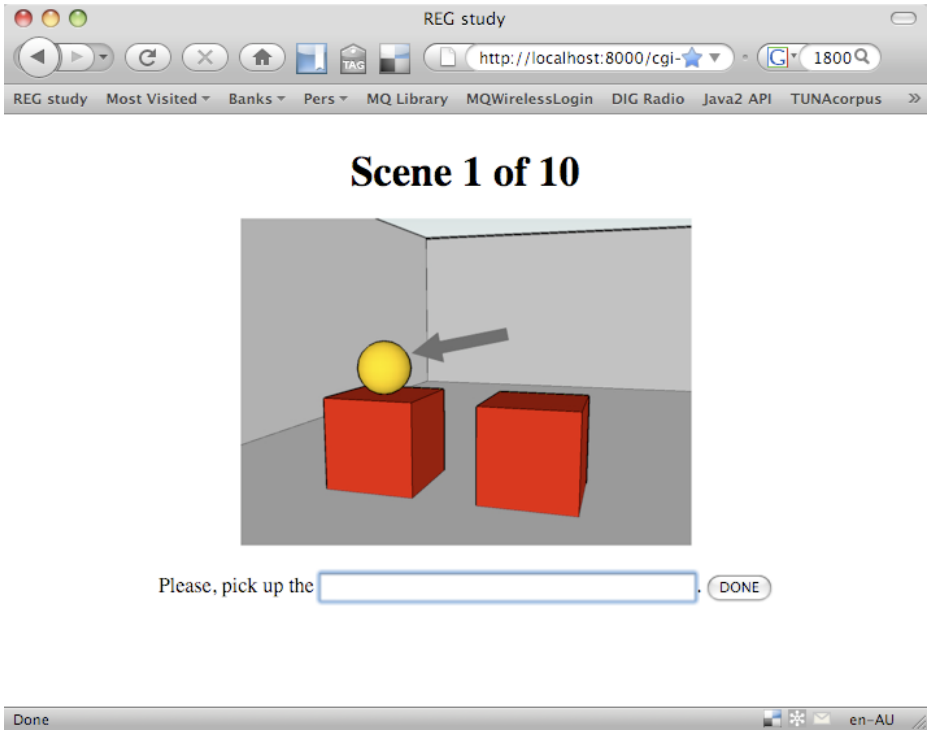

Fig. 1. An example stimulus scene.

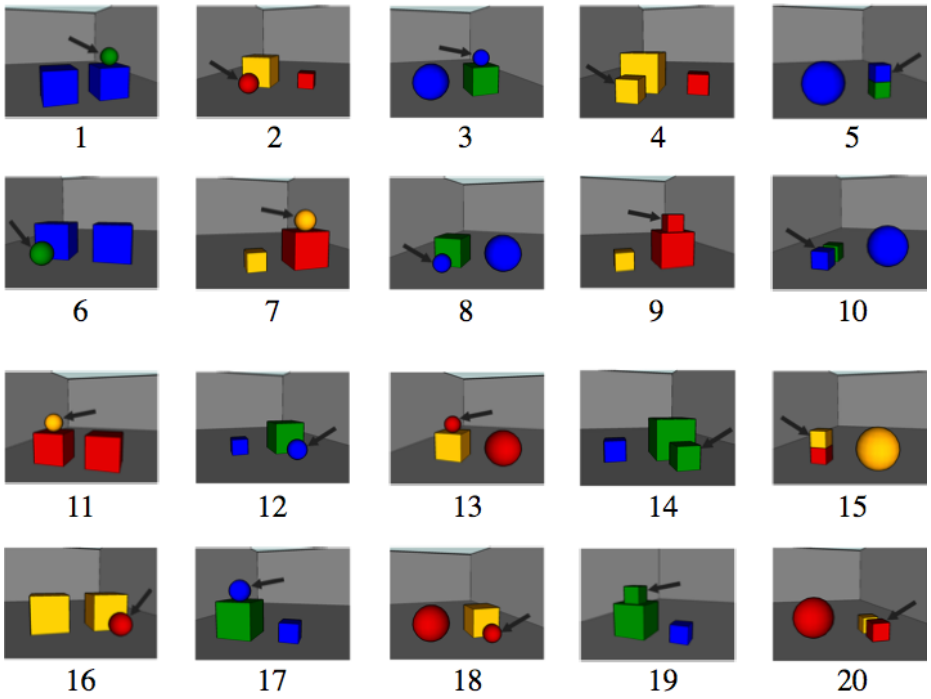

Fig. 2. The complete set of stimulus scenes.

Table 1. The 18 different patterns corresponding to the different forms of description that occur in the GRE3D3 corpus.

Label	Pattern	Example
A	⟨tg_col, tg_type⟩	*the blue cube*
B	⟨tg_col, tg_type, rel, lm_col, lm_type⟩	*the blue cube in front of the red ball*
C	⟨tg_col, tg_type, rel, lm_size, lm_col,lm_type⟩	*the blue cube in front of the large red ball*
D	⟨tg_col, tg_type, rel, lm_size, lm_type⟩	*the blue cube in front of the large ball*
E	⟨tg_col, tg_type, rel, lm_type⟩	*the blue cube in front of the ball*
F	⟨tg_size, tg_col, tg_type⟩	*the large blue cube*
G	⟨tg_size, tg_col, tg_type, rel, lm_col,lm_type⟩	*the large blue cube in front of the red ball*
H	⟨tg_size, tg_col, tg_type, rel, lm_size, lm_col, lm_type⟩	*the large blue cube in front of the large red ball*
I	⟨tg_size, tg_col, tg_type, rel, lm_size, lm_type⟩	*the large blue cube in front of the large ball*
J	⟨tg_size, tg_col, tg_type, rel, lm_type⟩	*the large blue cube in front of the ball*
K	⟨tg_size, tg_type⟩	*the large cube*
L	⟨tg_size, tg_type, rel, lm_size, lm_type⟩	*the large cube in front of the large ball*
M	⟨tg_size, tg_type, rel, lm_type⟩	*the large cube in front of the ball*
N	⟨tg_type⟩	*the cube*
O	⟨tg_type, rel, lm_col, lm_type⟩	*the cube in front of the red ball*
P	⟨tg_type, rel, lm_size, lm_col, lm_type⟩	*the cube in front of the large red ball*
Q	⟨tg_type, rel, lm_size, lm_type⟩	*the cube in front of the large ball*
R	⟨tg_type, rel, lm_type⟩	*the cube in front of the ball*

required to type fully distinguishing referring expressions for each trial. For the research presented here we also removed 7 descriptions which failed to uniquely distinguish the target referent from the other two objects.

The resulting corpus consists of 623 descriptions. As we are at this point only interested in the semantic content of these descriptions, we carried out a rigourous lexical and syntactic normalisation. In particular, we corrected spelling mistakes; normalised names for *colour* values and head nouns (such as *box* instead of *cube*); and replaced complex syntactic structures such as relative clauses with semantically equivalent simpler ones such as adjectives.

This normalisation makes it apparent that every one of the descriptions is an instance of one of the 18 content patterns shown in Table 1; for ease of reference, we label these patterns A through R. Each pattern indicates the set of attributes used in the description, where each attribute is identified by the object it describes (tg for target, lm for landmark) and the attribute used (col, size and type for colour, size and type, respectively).

Table 2. The number of occurrences of each description pattern across the ten stimulus scenes. Counts of 5 or more are shown in bold to make it easier to see the patterns, on the basis that low counts might be considered noise; of course, 5 is a fairly arbitrary cut-off point.

	Scene #									
Pattern	1	2	3	4	5	6	7	8	9	10
A	**17**	**23**			**37**	**38**	**25**			**39**
B	**14**	**8**	3		**16**	**7**	**8**	3		**10**
C		4		1			3		1	
D						1	1		1	
E	4		1			2				
F	1	1	**14**	**43**	4	3	2	**24**	**39**	**8**
G	1		**14**		2		1	**14**		1
H		1	1	**13**	2	1	2	1	**17**	2
I				3					1	
L				1						
M	1		**7**					4		
N	**11**	**13**				**7**	**14**			
O		4					1			
P							1			
Q		3					2			
R	**13**	**5**	**9**			2	2	1		

The description patterns in Table 1 are collected from across all the stimuli scenes; but of course not all patterns are equally common for every scene. Table 2 shows the distribution of the different patterns across the different scenes. We collapse the pairwise equivalent scenes from the two stimulus sets into one, so

that we are now only dealing with ten scenes. The primary observation of relevance here is that there is no one correct answer for how to refer to the target in any given scene: for example, Scenes 4, 5, 9 and 10 result in five semantically distinct referring expression forms, and Scene 7 results in 12 distinct referring expression forms. All of these are distinguishing descriptions, so all are acceptable forms of reference, although some contain more redundancy than others. Most obvious from the table is that, for many scenes, there is a predominant form of reference used; for example, pattern F ($\langle \mathsf{tg\_size}, \mathsf{tg\_col}, \mathsf{tg\_type} \rangle$) accounts for 43 (68%) of the descriptions used in Scene 4, and pattern A ($\langle \mathsf{tg\_col}, \mathsf{tg\_type} \rangle$) is very frequently used in a number of scenes.

Faced with such variation, the question arises: what exactly should we be trying to model when we develop algorithms for the generation of referring expressions? Clearly we could try to model an individual speaker, although this would require considerably more data than we have available here, and it remains unclear just what the utility of such a modelling exercise would be; we would also need to have some way of accounting for the fact that speakers are not necessarily consistent with themselves. An alternative would be to try to determine what characteristics of human referential behaviour, whether widely-used or specific to a select few 'good referrers', are particularly effective from the hearer's point of view, and then to try to model such an 'optimal' generation strategy.

Ultimately, practical applications of NLG techniques are more likely to be able to make use of algorithms developed from the latter perspective. To get to such a point, we have to first dissect referential behaviour in order to determine what kinds of characteristics might be relevant; the aim of this chapter is to reframe the referring expression generation problem in such a way as to shed light on the processes involved.

3 An Alternative Paradigm

Given an intended referent R, a set of distractors C, a set of attributes L_R, and the set of attributes to use in a description D:

```
1   Let D = ∅
2   repeat
3       add a selected attribute ∈ L_R to D
4       recompute C given D
5   until C = ∅
```

Fig. 3. The structure of referring expression generation algorithms.

The Incremental Algorithm and a considerable number of other algorithms proposed in the literature share the structure shown in Figure 3. Starting with

an empty set of properties D to be used in the referring expression and a set of distractors C, they add one property at a time to D (Line 3), and check after each step whether all distractors from C are ruled out yet (Lines 4 and 5).

What makes one algorithm different from another, in terms of this schema, is the particular means by which the next attribute to use is selected (Line 3, the first step inside the repeat loop): in the Greedy Algorithm [1], for example, the most useful attribute (i.e., that which rules out most potential distractors) is chosen next, whereas in the Incremental Algorithm the next attribute chosen is selected according to a pre-determined preference order.

In earlier work [17], we observed that the kinds of variation found in a particular set of human data might be modelled by using different preference orderings in the Incremental Algorithm. But this requires not only that potentially different preference orders are used by different speakers, but even that different preference orders are used by the same speaker on different occasions. Constantly switching preference orders does not seem to us to be a very convincing explanation for what humans do. More fundamentally, our view is that the schema shown in Figure 3 does not seem very convincing as a characterisation of the cognitive processes involved in producing referring expressions, requiring as it does a check of the scene (the second step in the repeat loop) after each attribute is added in order to determine whether further work should be done. There may be contexts where such a careful strategy is pursued, but it seems less likely that this is what people do each and every time they construct a referring expression. We do not, therefore, consider the Incremental Algorithm or its derivatives (or any other algorithms that share the schematic structure shown in Figure 3) to be convincing models of human behaviour.

One element of the Incremental Algorithm that we do find appealing, however, is its notion that there might be a 'force of habit' element to referential behaviour, as encoded in the preference ordering: for example, a preference order which proposes that an object's colour should be the first candidate attribute to be tried is just a way of saying that, for this speaker (or the speaker represented by such a preference ordering), 'colour is often useful so we'll try it first'. Clearly such a heuristic will not always be used—if all the objects in a scene are of the same colour, we might think it unlikely that a speaker would even consider the use of colour before ruling it out. However, in very many circumstances, it happens to be the case that colour is useful.

This leads us to hypothesise that there may be straightforwardly-apparent aspects of scenes that a speaker uses to determine what information is likely to be useful in producing a referring expression. For example, at least up to a certain number of objects in a scene, if the target referent is of one colour and all the other objects are of another, then, rather than a careful strategy of considering each attribute on its merits as embodied in conventional algorithms, it seems to us more plausible that some kind of gestalt perceptual strategy might be in play: the speaker knows, just by looking at the scene and without any algorithmic computation, that colour is a useful attribute for picking out the intended referent.

Of course, this is somewhat vague. To try to make the idea more concrete, we can look at the data available at a different grain-size than is usually done in algorithmic studies: rather than considering the overall content of the referring expressions produced by speakers, we can look at the individual attributes that occur in our data set. We can then establish how the gestalt perceptual characteristics of the scenes correspond to the use of each attribute, to see whether there are patterns across speakers that suggest more commonality than is apparent in the variety of data represented in Table 2.

4 What We Can Learn From the Data

4.1 Learning Algorithms for Description Construction

Table 3. The 10 characteristics of scenes.

Label	Attribute	Values
tg_type = lm_type	Target and Landmark share Type	TRUE, FALSE
tg_type = dr_type	Target and Distractor share Type	TRUE, FALSE
lm_type = dr_type	Landmark and Distractor share Type	TRUE, FALSE
tg_col = lm_col	Target and Landmark share Colour	TRUE, FALSE
tg_col = dr_col	Target and Distractor share Colour	TRUE, FALSE
lm_col = dr_col	Landmark and Distractor share Colour	TRUE, FALSE
tg_size = lm_size	Target and Landmark share Size	TRUE, FALSE
tg_size = dr_size	Target and Distractor share Size	TRUE, FALSE
lm_size = dr_size	Landmark and Distractor share Size	TRUE, FALSE
rel	Relation between Target and Landmark	on_top_of, in_front_of

It seems obvious that the visual context of reference must play at least some role in the choice of attributes in a given referring expression. An obvious question is then whether we can learn the description patterns in this data from the contexts in which they were produced. To explore this, we can capture the relevant aspects of context by means of a notion of *characteristics of scenes*. The characteristics of scenes which we hypothesise might have an impact on the choice of referential form in our data are those summarised in Table 3.[2] Each of these captures some aspect of the scene which we consider to be readily apparent without complex computation on the part of the perceiver, a point we will return to later. Each feature compares two of the three objects in the scene to each other with respect to their values for one of the attributes type, colour and size.

[2] Note that the spatial relations between the distractor and the other two objects are not listed as characteristics of the scenes because they are the same in all scenes. Whether the distractor is to the left or the right of the other two objects has no impact.

In this way we capture how common the landmark's and the target's properties are overall, which is a simple way of approximating their visual salience.

We used the implementation provided in the Weka toolkit [18] of the C4.5 decision tree classifier [19] to see what correspondences might be learned between these characterisics of scenes and the forms of referring expression shown in Table 1. The pruned decision tree learned by this method predicted the actual form of reference used in only 48% of cases under 10-fold cross-validation. On the basis of the discussion in the previous section, this is not surprising: Given that there are many 'gold standard' descriptions for each scene, a low score is to be expected. A mechanism which learns only one answer will inevitably be 'wrong'—in the sense of not replicating the human-produced description—in many cases, even if the description produced is still a valid distinguishing description.

More revealing, however, is the rule learned from the data:

if tg_type = dr_type
then use F ($\langle$tg_size, tg_col, tg_type$\rangle$)
else use A ($\langle$tg_col, tg_type$\rangle$)

Patterns A and F are the two most prevalent patterns in the data, and indeed one or the other appears at least once in the human data for each scene. If we analyse the learned rule in more detail, we see that it predicts Pattern F for Scenes 4, 5, 8, 9, 14, 15, 18, and 19, and A for all other scenes. The pattern distribution in Table 2 shows that this means that the rule is able to produce a 'correct' answer (in the sense of emulating what at least one speaker did) for every scene.

As there is clearly another factor at play causing variation in the data, we then re-ran the classifier, this time using the participant ID as well as the scene characteristics in Table 3 as features. This improved pattern prediction to 57.62%. This suggests that individual differences may indeed be capturable from the data, although we would need more data than the mere 10 examples we have from each participant to learn a good predictive model for a single speaker.

Table 4. Accuracy of learning attribute inclusion. Statistically significantly increases ($p<0.01$) are marked in bold.

Attribute to Include	Baseline (0-R)	Using Scene Characteristics	Using Scene Characteristics and Participant
tg_col	78.33%	78.33%	**89.57%**
tg_size	57.46%	**90.85%**	90.85%
rel	64.04%	65.00%	**81.22%**
lm_col	74.80%	**87.31%**	**93.74%**
lm_size	88.92%	**95.02%**	95.02%

4.2 Learning Heuristics for Attribute Inclusion

Attribute Inclusion Based on Scene Characteristics The experiments just described demonstrate that there is indeed considerable variation across speakers, and put into question any attempt to model human referring behaviour that ignores this. On the other hand, it seems implausible that there are no commonalities whatsoever between speakers. The alternative approach we propose here is to look for commonalities in the data in terms of the *constituent elements* of the different forms of reference used for each scene, rather than at the level of complete descriptions: are there characteristics of scenes which are highly likely to result in *specific attributes* being used in descriptions? This way of thinking about the data was foreshadowed in [16], where we observed that our participants could be separated into those who always used relations, those who never used relations, and those who sometimes used relations.

As a baseline here, we use the success rate of simply predicting the majority class. We might think of this as a 'context-free' approach, in the sense that the particular context of reference plays no role in the decision as to whether or not an attribute should be used. Table 4 compares the results for this approach with one model that is trained on the characteristics of scenes, and another that takes both the characteristics of scenes and the participant ID into account.[3]

The 'context-free' strategies work suprisingly well for predicting the inclusion of some attributes in the human data. As has been noted in other work, colour is often included in referring expressions irrespective of its discriminatory power, and this is borne out by the data here. Perhaps more suprising is the large degree to which the inclusion of landmark size is captured by a context-free strategy.

Improvement on all attributes other than target colour increases when we take into account the characteristics of the scenes, confirming the widely-held assumption that the context of reference does indeed make a difference. When we add participant ID to the features used in the learner, performance improves further still, indicating that there are speaker-specific consistencies across contexts. The numbers suggest that colour is effectively a participant property, whereas size is a scene property.

It is interesting to look at the rules learned on the basis of the scene characteristics alone; these are shown in Figure 4. Not surprisingly, the rule derived for target colour inclusion is simply to always include the colour (i.e., the same context-free colour inclusion rule that proves most effective in modelling the data without reference to scene characteristics). The target's size is included if target and distractor are of the same type (Scenes 2, 4, 7, 9, 12, 14, 17, 19); the spatial relation between the target and the landmark is included if the target is on top of the landmark and the landmark is of the same size as the distractor (Scenes 1, 3, 11, 13); the landmark's colour is included in all relational descriptions; and the landmark's size goes into all relational descriptions if the target and landmark cannot be distinguished by colour (Scenes 4, 9, 14 and 19).

[3] As before, the results reported are for the accuracy of a pruned decision tree, under 10-fold cross-validation.

Target Colour:
include tg_col

Target Size:
if tg_type = dr_type **then** include tg_size

Relation:
if rel = on_top_of and lm_size = dr_size **then** include rel

Landmark Colour:
if we have used a relation **then** include lm_col

Landmark Size:
if we have used a relation and tg_col = lm_col **then** include lm_size

Fig. 4. Rules learned on the basis of scene characteristics.

Attribute Inclusion on a Speaker-by-Speaker Basis The rules learned when we include participant ID are more complex, but can be summarised in a way that demonstrates how this approach can reveal something about the variety of ways in which speakers might be approaching the task of referring expression generation.

Focussing on the question of whether or not to use *the target object's colour* in a referring expression, the learner identifies five heuristics, which apply to the 63 participants as follows:

- For 37 participants it learned to always use colour, irrespective of the context (this corresponds to the baseline rule learned above).
- For the rest of the participants it always uses colour if the target and the landmark are of the same type (which again is intuitively quite appropriate).
- When the target and the landmark are not of the same type, we see more variation in learned behaviour:
  - for 19 participants colour simply doesn't get used;
  - in scenes where the target is on top of the landmark, six participants use colour if the target and the distractor have the same size, with two of these six always using colour in scenes where the target is in front of the landmark, and the other four using colour only if target and distractor do not share size; and
  - one participant is characterised as using colour if the target and distractor do not share colour.

Participants showed least commonality when it came to heuristics for the inclusion of a *relation*. For 15 participants, the model predicts that they always use a relation; 29 are predicted never to include a relation. Of the remaining participants, eight share a heuristic with one other participant, while the remaining 11 have a unique decision pattern. In total, 17 different rules were learned to account for the inclusion of relations in the descriptions.

Landmark colour is predicted to be used in all relational descriptions by 18 participants and in none by eight participants. For four participants, the use of the landmark's colour is predicted if the landmark shares its type with the target object; one of these four also includes the landmark colour for all cases where the target is in front of the landmark, and two participants use the landmark colour if target and distractor share size. The remaining seven participants each have their own heuristic. This results in 15 different heuristics.

As indicated by the lack of change in accuracy of prediction when participant ID is included (see Table 4), all participants are predicted to share the same heuristics for the inclusion of the *target's size* and the *landmark's size*, as shown in Figure 4 for these two properties. Size is therefore the property with the highest degree of commonality between the participants who contributed to this corpus.

The specific content of the rules mentioned above may appear idiosyncratic; they are just what the limited data in the corpus supports, and some elements of the rules may be due to artefacts of the specific stimuli used in the data gathering. We would require a more diverse set of stimuli to determine whether this is the case, but the basic point stands: *we can find correlations between characteristics of the scenes and the presence or absence of a particular attribute in referring expressions, even if we cannot predict so well the exact combinations of these correlations that a given speaker will use*. Of course, this is in some sense what all referring expression generation algorithms aim to capture; our claim here is that an attribute-centric model is more able to explain the human data.

From the behaviour of the 63 participants that contributed to the data set used in this study, we learn 30 different combinations of these attribute-inclusion heuristics. One way to think of this is that each combination of attribute-inclusion heuristics corresponds to a *speaker profile*. We have avoided listing the details of all 30 speaker profiles here; however, the sets of heuristics that were each used by more than one participant are shown in Table 5.

So, for example, our data contains 13 people who automatically include the colour of an intended referent, never use a relation to a landmark, and make the inclusion of the referent's size dependent on the similarity of the referent to another object. We also have two people who make the inclusion of all attributes dependent on contextual factors, with the exception of the landmark's colour, which they never mention; and so on. Grouping people according to the collection of individual attribute-heuristics they use can be seen as a more fine-grained alternative to predicting a particular content pattern for each participant–scene combination, which also more easily generalises to new scenes.

5 What's Missing?

As we have argued above, and as captured schematically in Figure 3, existing approaches to the generation of referring expressions are primarily focussed on the development of algorithms whose main function is to control the serial incorporation of attribute values into a developing referring expression. This focus has

Table 5. The most common speaker profiles. The first column indicates the number of speakers sharing each profile.

#	tg_col	tg_size	tg_size	rel	lm_size
13	TgCol-T	TgSize-1	Rel-F	n/a	n/a
10	TgCol-T	TgSize-1	Rel-T	LmCol-T	LmSize-1
9	TgCol-1	TgSize-1	Rel-F	n/a	n/a
2	TgCol-3	TgSize-1	Rel-4	LmCol-F	LmSize-1
2	TgCol-T	TgSize-1	Rel-2	LmCol-T	LmSize-1
2	TgCol-1	TgSize-1	Rel-T	LmCol-1	LmSize-1

TgCol-T = always include tg colour
TgCol-1 = include tg colour if tg and lm share type.
TgCol-3 = include tg colour if tg and lm share type
or if the tg is in front of the lm
or if tg and dr share size.
TgSize-1 = include tg size if tg and dr share type.
Rel-F = never include a relation.
Rel-T = always include a relation.
Rel-2 = include a relation if tg shares type with dr but not lm.
Rel-4 = include a relation if the tg is on top of the lm
and lm and dr share size
LmCol-T = if a relation is present, include the lm colour.
LmCol-F = never include lm colour.
LmCol-1 = include lm colour if tg and lm share type.
LmSize-1 = if a relation is present, include the lm size
if tg and lm share colour.

either meant that all attributes are considered a priori equal (as in algorithms which compute discriminatory power, and use this as the sole determinant for attribute selection), or are ordered on the basis of some pre-defined but unexplained preferences (as in the Incremental Algorithm and its variants, as well as Krahmer et al.'s [9] graph-based algorithm).

We have proposed here a rather more bottom-up way of thinking about referring expression generation: rather than focus on algorithms whose purpose is to control the combination of individual attributes, we argue instead that we should focus on the individual attributes themselves, and explore what it is that makes them appropriate for inclusion in a developing referring expression in a given situation. The picture that emerges is one where we can think of an individual speaker's approach to reference as consisting of a collection of attribute-specific heuristics. These individual heuristics are shared across speakers to a greater or lesser degree, and the combinations of heuristics vary by speaker. A speaker may even use different heuristics depending upon features orthogonal to the referential context (such as who they are talking to, or the mission-critical nature of getting it right first time).

As we have suggested above, this makes for a much richer story of how humans produce referring expressions, and provides a basis for the development of algorithms that enable us to incorporate much more sophistication: depending on the circumstances, a referring expression might be generated almost without thinking, simply on the basis of general properties of a scene; or, in mission-critical situations, a more cautious, reflective and reasoned approach might be used.[4]

Clearly the model we have sketched here is not a complete picture of how referring expressions might be generated. What we have argued is that, at least in simple scenes, effective referring expressions almost 'jump out' at the speaker, without any need for complex reasoning or retrospective analysis to determine whether they are indeed successful distinguishing descriptions. Particulary as scenes get more complex, it is perhaps unlikely that referring expressions which just happen to be distinguishing can be 'read off' the scene, and more plausible that some reasoning is required by the hearer in order to determine whether the expression constructed so far is sufficient for the identification task at hand. Van der Sluis [21] cites an example of a referring expression that appears in an episode of the television series *Twin Peaks*, where a character by the name of Lucy is attempting to identify a referent for a speaker:

> Uhm, I'm gonna transfer to the phone on the table by the red chair ... [points in the direction of the phone] the ... the red chair, against the wall, uh the little table, with the lamp on it, the lamp that we moved from the corner? ... the black phone, not the brown phone ...

Examples like this suggest that, in appropriately complex referential situations, what we have is a repeated iteration between 'reading off' attributes from the

[4] The distinction here is deliberately reminiscent of Carletta's risky vs cautious distinction [20] in regard to referential behaviour in the HCRC Map Task Corpus.

scene, and reflective analysis that indicates more work is needed to achieve successful reference; however, unlike existing algorithms, there is no need for the success check to be carried out after each and every attribute. To model this kind of behaviour, we need to develop algorithms which, although they may construct a first-pass referring expression by means of a kind of 'parallel gestalt', are then open to a process of selective extension whereby a monitoring process (somewhat akin to the anticipation-feedback loop of [22]) decides what needs to be added in a more reasoned way. Our existing data does not allow us to explore these kinds of questions, and we are not aware of any data sets that do; but the framework we have sketched is suggestive of experiments that might be carried out to probe these kinds of phenomena more delicately.

6 Conclusions

In this chapter, we have suggested that the bulk of the existing work on referring expression generation, including our own previous work, has mistakenly placed the focus on the development of algorithms for combining attributes in a serial fashion to produce distingishing descriptions. We have suggested an alternative way of thinking about the problem, where the focus is instead placed on the specific attributes that make up a referring expression. By doing this, we have shown how it is possible, despite the apparent wide variation in the forms of reference that speakers produce, to identify a number of component strategies that are common across many speakers. Under this analysis, the apparently broad variation that we see is a result of different speakers using different combinations of component strategies which themselves vary in the extent to which they are shared across speakers.

There are many questions left unresolved here. As we suggested earlier, the practical application of NLG techniques for generating referring expressions—as might be required in object location in 'omniscient room' scenarios or landmark description in navigation systems—is probably best served by identifying the component strategies that are most effective, and we have not touched on that question here. Also, as discussed in Section 5, the picture we have presented here is only a part of the story, and a complete algorithm still requires the integration of these component strategies into a model which is capable of extending a referring expression appropriately when it is inadequate for the task. However, we believe the framework described here provides a new perspective on what is involved in referring expression generation, and points the way to a range of experimental studies that will provide greater insight.

References

1. Dale, R.: Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, Vancouver BC, Canada (1989)

2. Dale, R., Haddock, N.: Content determination in the generation of referring expressions. Computational Intelligence **7**(4) (1991) 252–265
3. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science **19**(2) (1995) 233–263
4. van Deemter, K.: Generating referring expressions: Boolean extensions of the Incremental Algorithm. Computational Linguistics **28**(1) (2002) 37–52
5. Gardent, C.: Generating minimal definite descriptions. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia PA, USA (2002)
6. Horacek, H.: On referring to sets of objects naturally. In: Proceedings of the 3rd International Conference on Natural Language Generation, Brockenhurst, UK (2004) 70–79
7. Jordan, P.W.: Contextual influences on attribute selection for repeated descriptions. In van Deemter, K., Kibble, R., eds.: Information Sharing: Reference and Presupposition in Language Generation and Interpretation. CSLI Publications, Stanford CA, USA (2002)
8. Krahmer, E., Theune, M.: Efficient context-sensitive generation of referring expressions. In van Deemter, K., Kibble, R., eds.: Information Sharing: Reference and Presupposition in Language Generation and Interpretation. CSLI Publications, Stanford CA, USA (2002) 223–264
9. Krahmer, E., van Erk, S., Verleg, A.: Graph-based generation of referring expressions. Computational Lingustics **29**(1) (2003) 53–72
10. van Deemter, K.: Generating referring expressions that involve gradable properties. Computational Linguistics **32**(2) (2006) 195–222
11. Gatt, A., van Deemter, K.: Conceptual coherence in the generation of referring expressions. In: Proceedings of the 21st COLING and the 44th ACL Conference, Sydney, Australia (2006)
12. Kelleher, J., Kruijff, G.J.M.: Incremental generation of spatial referring expressions in situated dialog. In: Proceedings of the 21st COLING and the 44th ACL Conference, Sydney, Australia (2006)
13. Belz, A., Kow, E., Viethen, J., Gatt, A.: The GREC challenge 2008: Overview and evaluation results. In: Proceedings of the 5th International Natural Language Generation Conference, Salt Fork OH, USA (2008) 183–191
14. Gatt, A., Belz, A., Kow, E.: The TUNA challenge 2008: Overview and evaluation results. In: Proceedings of the 5th International Natural Language Generation Conference, Salt Fork OH, USA (2008) 198–206
15. Viethen, J., Dale, R.: Generating referring expressions: What makes a difference? In: Australasian Language Technology Association Workshop 2008, Hobart, Australia (2008) 160—168
16. Viethen, J., Dale, R.: The use of spatial relations in referring expression generation. In: Proceedings of the 5th International Conference on Natural Language Generation, Salt Fork OH, USA (2008)
17. Viethen, J., Dale, R.: Algorithms for generating referring expressions: Do they do what people do? In: Proceedings of the 4th International Conference on Natural Language Generation, Sydney, Australia (2006) 63–70
18. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco CA, USA (2005)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco CA, USA (1993)
20. Carletta, J.C.: Risk-taking and Recovery in Task-Oriented Dialogue. PhD thesis, University of Edinburgh (1992)

21. van der Sluis, I.: Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions. PhD thesis, Tilburg University, The Netherlands (2005)
22. Jameson, A., Wahlster, W.: User modelling in anaphora generation: ellipsis and definite description. In: Proceedings of the 5th European Conference on Artificial Intelligence, Orsay, France (1982) 222–227