Referring Expression Generation through Attribute-Based Heuristics

Robert Dale and Jette Viethen
Centre for Language Technology
Macquarie University
Sydney, Australia
rdale@ics.mq.edu.au|jviethen@ics.mq.edu.au

Abstract

In this paper, we explore a corpus of human-produced referring expressions to see to what extent we can learn the referential behaviour the corpus represents. Despite a wide variation in the way subjects refer across a set of ten stimuli, we demonstrate that component elements of the referring expression generation process appear to generalise across participants to a significant degree. This leads us to propose an alternative way of thinking of referring expression generation, where each attribute in a description is provided by a separate heuristic.

1 Introduction

The last few years have witnessed a considerable move towards empiricism in referring expression generation; this is evidenced both by the growing body of work that analyses and tries to replicate the content of corpora of human-produced referring expressions, and particularly by the significant participation in the TUNA and GREC challenge tasks built around such activities (see, for example, (Belz and Gatt, 2007; Belz et al., 2008; Gatt et al., 2008)). One increasingly widespread observation—obvious in hindsight, but surprisingly absent from much earlier work on referring expression generation—is that one person's referential behaviour differs from that of another: given the same referential task, different subjects will choose different referring expressions to identify a target referent. Faced with this apparent lack of cross-speaker consistency in how to refer to entities, we might question the validity of any exercise that tries to develop an algorithm on the basis of data from multiple speakers.

In this paper we revisit the corpus of data that was introduced and discussed in (Viethen and Dale, 2008a; Viethen and Dale, 2008b) with the objective of determining what referential behaviour, if any, might be learned automatically from the data. We find that, despite the apparent diversity of the data when we consider the production of referring expressions across subjects, a closer examination reveals that individual attributes within referring expressions do appear to be selected on the basis of contextual factors with a high degree of consistency. This suggests that referring behaviour might be best thought of as consisting of a combination of lower-level heuristics, with each individual's overall referring behaviour being constructed from a potentially distinct combination of these common heuristics.

In Section 2 we describe the corpus we use for the experiments in this paper. In Section 3 we explore to what extent we can use this corpus to learn an algorithm for referring expression generation; in Section 4 we look more closely at the nature of individual variation within the corpus. Section 5 briefly discusses related work on the use of machine learning in referring expression generation, and Section 6 draws some conclusions and points to future work.

2 The Corpus

2.1 General Overview

The corpus we use was collected via a data gathering experiment described in (Viethen and Dale, 2008a; Viethen and Dale, 2008b). The purpose of the data gathering was to gain some insight into how human subjects use relational referring expressions, a relatively unexplored aspect of referring expression generation. Participants visited a website, where they first saw an introductory page with a set of simple instructions and a sample stimulus scene consisting of three objects. Each participant was then assigned one of two trial sets of ten scenes each; the two trial sets are superficially

Proceedings of the 12th European Workshop on Natural Language Generation, pages 58–65,
Athens, Greece, 30 – 31 March 2009.

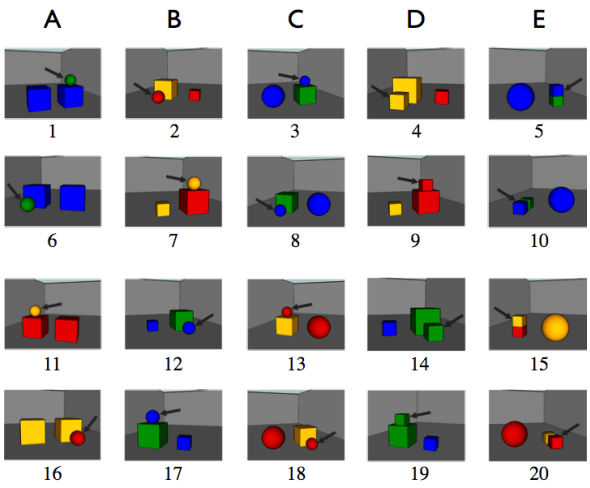

Figure 1: The stimulus scenes. The letters indicate which schema from Figure 2 each column of scenes is based on.

different, 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 1, where Trial Set 1 consists of Scenes 1 through 10, and Trial Set 2 consists of Scenes 11 through 20.[1]

The scenes were presented successively in a preset order, which was the same for each participant. Below each scene, the participant had to complete the sentence *Please pick up the* ... in a text box before clicking on a button to see the next 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.

The experiment was completed by 74 participants from a variety of different backgrounds and ages; most were university-educated and in their early or mid twenties. For reasons discussed in (Viethen and Dale, 2008b), the data of 11 participants was discarded. Of the remaining 63 participants, 29 were female, while 34 were male.

2.2 Stimulus Design

The design of the stimuli used in the experiment is described in detail in (Viethen and Dale, 2008a).

[1]Scene 1 is paired with Scene 11, Scene 2 with Scene 12, and so on; in each pair, the only differences are the colour scheme used and the left–right orientation, with these variations being introduced to make the experiment less monotonous for subjects; (Viethen and Dale, 2008a) report that these characteristics of the scenes appear to have no significant effect on the forms of reference used.

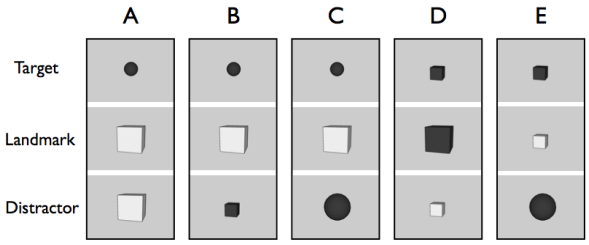

Figure 2: The *schemata* which form the basis for the stimulus scenes.

We provide a summary of the key points here.

In order to explore even the most basic hypotheses with respect to the use of relational expressions, which was the aim of the original study, scenes containing at least three objects were required. One of these objects is the intended referent, which is referred to here as the *target*. The subject has to describe the target in such a way as to distinguish it from the other two objects in the scene. Although the scenes presented to the subjects are such that spatial relations are never *necessary* to distinguish the target, they are set up so that one of the two non-target objects was clearly closer to the target. This object is referred to as the (potential) *landmark*; and we call the third object in the scene the *distractor*.

To minimise the number of variables in the experiments, scenes are restricted to only two kinds of objects, cubes and balls. The objects also vary in two dimensions: colour (either green, blue, yellow, or red); and size (either large or small).

To further reduce the number of factors in the scene design, the landmark and distractor are always placed clearly side by side, and the target is located on_top_of or directly in_front_of the landmark.

Finally, to reduce the set of possible stimuli to a manageable number, five *schemata* (see Figure 2) were created as a basis for the final stimulus set. The design of these schemata was informed by a number of research questions with regard to the use of relations; see (Viethen and Dale, 2008b). A schema determines the type and size of each object in the scenes that are based on it, and determines which objects share colour. So, for example, in scenes based on Schema C, the target is a small ball; the landmark is a large cube with different colour from the target; and the distractor is a large ball sharing its colour with the target.

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*

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

From each schema, four distinct scenes were generated, resulting in the 20 stimulus scenes shown in Figure 1. As noted above, there are really only 10 distinct 'underlying' scene types here, so in the remainder of this paper we will talk in terms of Scenes 1 through 10, where the data from the pairwise matched scenes are conflated.

2.3 The GRE3D3 Corpus[2]

Before conducting any quantitative data analysis, some syntactic and lexical normalisation was carried out on the data provided by the participants. In particular, spelling mistakes were corrected; normalised names were used for colour values and head nouns (for example, *box* was replaced by *cube*); and complex syntactic structures such as relative clauses were replaced with semantically equivalent simpler ones such as adjectives. These normalisation steps should be of no consequence to the analysis presented here, since we are solely interested in exploring the *semantic* content of referring expressions, not their lexical and syntactic surface structure.

For the purposes of the machine learning experiments described in this paper, we made a few further changes to the data set in order to keep the number of properties and their possible values low. We removed locative expressions that made reference to a part of the scene (58 instances) and references to size as *the same* (4 instances); so, for example, *the blue cube on top of the green cube in the right* and *the blue cube on top of the green cube of the same size* both became *the blue cube on top of the green cube*. We also removed the mention of a third object from ten descriptions in order to keep the number of possible objects per description to a maximum of two. These changes resulted in seven descriptions no longer satisfying the criterion of being fully distinguishing, so we removed these descriptions from the corpus.

3 Learning Description Patterns

The resulting corpus consists of 623 descriptions. Every one of these is an instance of one of the 18 patterns shown in Table 1; for ease of reference, we label these patterns A through R. Each pattern indicates the sequence 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).

Most work on referring expression generation attempts to determine what attributes should be used in a description by taking account of aspects of the context of reference. 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 chose to capture the relevant aspects of context by means of the notion of *characteristics of scenes*. The char-

[2]The data set resulting from the experiment described above is known as the GRE3D3 Corpus; the name stands for '**G**eneration of **R**eferring **E**xpressions in **3D** scenes with **3** Objects'.

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

Table 2: The 10 characteristics of scenes

acteristics of scenes which we hypothesize might have an impact on the choice of referential form are those summarised in Table 2; these are precisely the characteristics that were manipulated in the design of the schemata in Figure 2.

Of course, there is no one correct answer for how to refer to the target in any given scene. Figure 3 shows the distribution of different patterns across the different scenes; so, for example, some scenes (Scenes 4, 5, 9 and 10) result in only five semantically distinct referring expression forms, whereas 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 chart is that, for many scenes, there is a predominant form of reference used; so, for example, pattern F ($\langle$tg_size, tg_col, tg_type$\rangle$) accounts for 43 (68%) of the descriptions used in Scene 4, and pattern A ($\langle$tg_col, tg_type$\rangle$) is very frequently used in a number of scenes.[3]

We used Weka (Witten and Eibe, 2005) with the J48 decision tree classifer to see what correspondences might be learned between the characterisics of the scenes listed in Table 2 and the forms of referring expression used for the target referents, as 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, but given that there are many 'gold standard' descriptions for each scene, this low score is hardly surprising; a mechanism which learns only one answer will inevitably be 'wrong' in many cases. 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$\rangle$)
endif

Patterns A and F are the two most prevalent patterns in the data, and indeed one or other appears at least once in the human data for each scene; consequently, the learned rule is able to produce a 'correct' answer for every scene.[4]

4 Individual Variation

One of the most striking things about the data in this corpus is the extent to which different subjects appear to do different things when they construct referring expressions, as demonstrated by the distribution of patterns in Figure 3. Another way of looking at this variation is to characterise the behaviour of each subject in terms of the sequence of descriptions they provide in response to the set of 10 stimuli.

Across the 63 subjects, there are 47 different sequences; of these, only four occur more than once (in other words, 43 subjects did not produce the same sequence of descriptions for the ten scenes as anyone else). The recurrent sequences, i.e. those used by at least two people, are shown in Table 3. Note that the most frequently recurring sequence,

[3]The chart as presented here is obviously too small to enable detailed examination, and our use of colour coding will be of no value in a monchrome rendering of the paper; however, the overall shape of the data is sufficient to demonstrate the points we make here.

[4]The fact that the rule is conditioned on a property of the distractor object may be an artefact of the stimulus set construction; this would require a more diverse set of scenes to determine.

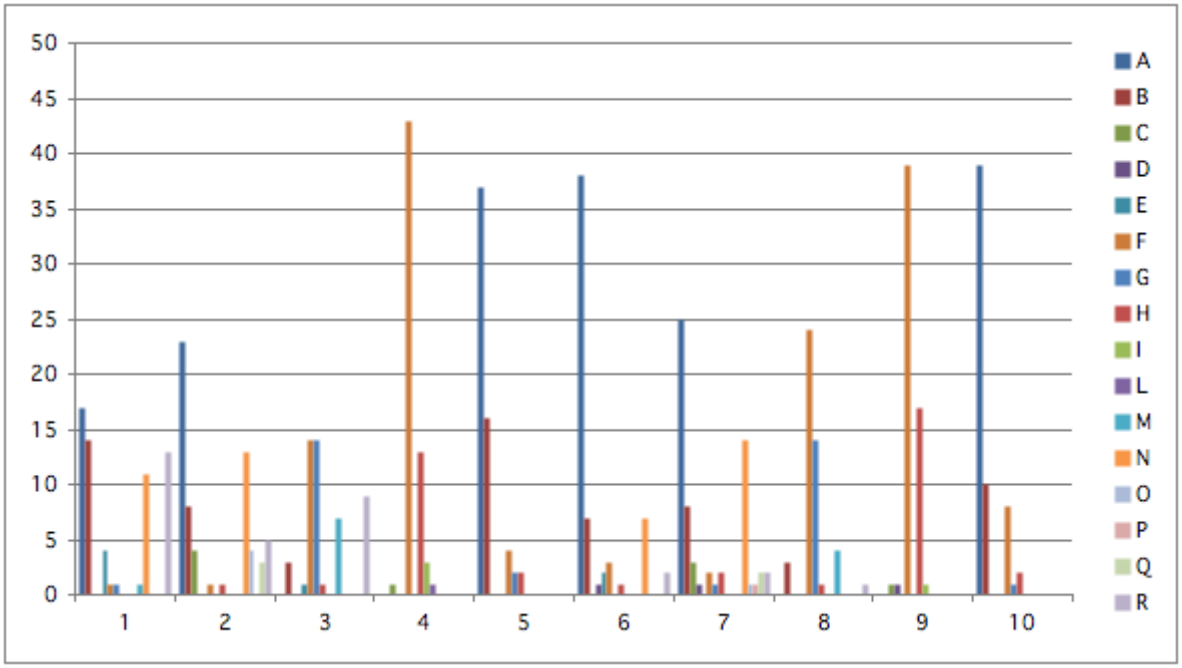

Figure 3: The profile of different description patterns (A through R) for each of the 10 scenes. The length of the bar indicates how often each of the 18 patterns is used.

which matches the behaviour of nine separate subjects, consists only of uses of patterns A and F. It remains to be seen to what extent a larger data set would demonstrate more convergence; however, the point to be made at present is that any attempt to predict the behaviour of a given speaker by means of a model of referring behaviour is going to have to take account of a great deal of indiviual variation.

Nonetheless, we re-ran the J48 classifier described in the previous section, this time using the participant ID as well as the scene characteristics in Table 2 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 subject to learn a good predictive model.

In the face of this lack of data, another approach is to look for commonalities in the data in terms of the *constituent elements* of the different reference patterns used for each scene. This way of thinking about the data was foreshadowed by (Viethen and Dale, 2008b), who observed that the subjects could be separated into those who always used relations, those who never used relations, and those who sometimes used relations. This leads us to consider whether there are characteristics of scenes or speakers which are highly likely to result in *specific attributes* being used in descriptions. If this is the case, a decision tree learner should be able to learn for each individual attribute whether it should be included in a given situation.

An appropriate baseline for any experiments here is the success rate of simply including or not including each attribute (basically a 0-R majority class classifier), irrespective of the characteristics of the scene. Table 4 compares the results for this 'context-free' 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.[5]

Interestingly, 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 (see for example (Viethen et al., 2008)), 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.

[5] As before, the results reported are for the accuracy of a pruned J48 decision tree, under 10-fold cross-validation.

Improvement on all attribues other than target colour improves when we take into account the characteristics of the scenes, consistent with our assumptions that context does matter. When we add participant ID to the features used in the learner, performance improves further still, indicating that there are speaker-specific consistencies in the data.

It is instructive to look at the rules learned on the basis of the scene characteristics. 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 rules for including the other attributes on the basis of scene characteristics (but not participant ID) are shown in Figure 4.

The rules learned when we include participant ID are more conplex, but can be summarised in a way that demonstrates how this approach can reveal something about the variety of ways in which speakers appear to approach the task of referring expression generation. Focussing, as an example, just on the question of whether or not to use the target object's colour in a referring expression, we find the following:

- 48 participants always used colour, irrespective of the context (this corresponds to the baseline rule learned above).
- The other participants always use colour if the target and the landmark are of the same type (which again is intuitively quite appropriate).
- When the landmark and the target are not of the same type, we see more variation in behaviour; 19 participants simply don't use colour, and the behaviour of seven can be captured via a more complex analysis: four use colour if the target and the distractor are the same size, two use colour if the target and distractor are of the same size and the target is on top of the landmark, and one uses colour if the target and distractor share colour.

Again, the specific details of the rules learned here are probably not particularly significant, based as they are on a limited data set and a set of stimuli that may give elevated status to incidental properties. However, the general point remains that we

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

Figure 4: Rules learned on the basis of scene characteristics

can use this kind of analysis to identify possible rules for the inclusion of individual attributes in referring expressions.

What this suggests is that we might be able to capture the behaviour of individual speakers not in terms of an overall strategy, but as a composite of heuristics, where each heuristic accounts for the inclusion of a specific attribute. The rules, or heuristics, shown in Figure 4 are just those which are most successful in predicting the data; but there can be many other rules that might be used for the inclusion of particular attributes. So, for example, I might be the kind of speaker who just automatically includes the colour of an intended referent without any analysis of the scene; and I might be the kind of speaker who always uses a relation to a nearby landmark in describing the intended referent. Or I might be the kind of speaker who surveys the scene and takes note of whether the landmark's colour is distinctive; and so on.

Thought of in this way, each speaker's approach to reference is like a set of 'parallel gestalts' that contribute information to the description being constructed. The particular rules for inclusion that any speaker uses might vary depending on their personal past history, and perhaps even on the basis of situation-specific factors that on a given occasion might lean the speaker towards either being 'risky' or 'cautious' (Carletta, 1992).

As alluded to earlier, the specific content of the rules shown in Figure 4 may appear idiosyncratic; they are just what the limited data in the corpus

Pattern Sequence ($\langle$Scene#, DescriptionPattern$\rangle$)	Number of subjects
$\langle 1, A\rangle$, $\langle 2, A\rangle$, $\langle 3, G\rangle$, $\langle 4, F\rangle$, $\langle 5, A\rangle$, $\langle 6, A\rangle$, $\langle 7, A\rangle$, $\langle 8, G\rangle$, $\langle 9, F\rangle$, $\langle 10, A\rangle$	2
$\langle 1, B\rangle$, $\langle 2, B\rangle$, $\langle 3, G\rangle$, $\langle 4, H\rangle$, $\langle 5, B\rangle$, $\langle 6, B\rangle$, $\langle 7, B\rangle$, $\langle 8, G\rangle$, $\langle 9, H\rangle$, $\langle 10, B\rangle$	2
$\langle 1, N\rangle$, $\langle 2, N\rangle$, $\langle 3, K\rangle$, $\langle 4, F\rangle$, $\langle 5, A\rangle$, $\langle 6, N\rangle$, $\langle 7, N\rangle$, $\langle 8, K\rangle$, $\langle 9, F\rangle$, $\langle 10, A\rangle$	6
$\langle 1, A\rangle$, $\langle 2, A\rangle$, $\langle 3, F\rangle$, $\langle 4, F\rangle$, $\langle 5, A\rangle$, $\langle 6, A\rangle$, $\langle 7, A\rangle$, $\langle 8, F\rangle$, $\langle 9, F\rangle$, $\langle 10, A\rangle$	9

Table 3: Sequences of description patterns found more than once

Attribute to Include	Baseline (0-R)	Using Scene Characteristics	Using Scene Characteristics and Participant
Target Colour	78.33%	78.33%	**89.57%**
Target Size	57.46%	**90.85%**	90.85%
Relation	64.04%	65.00%	**81.22%**
Landmark Colour	74.80%	**87.31%**	**93.74%**
Landmark Size	88.92%	**95.02%**	95.02%

Table 4: Accuracy of Learning Attribute Inclusion; statistically significant increases ($p<.01$) in bold.

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 particular attributes in referring expressions, even if we cannot predict so well the particular combinations of these correlations that a given speaker will use in a given situation.

5 Related Work

There is a significant body of work on the use of machine learning in referring expression generation, although typically focussed on aspects of the problem that are distinct from those considered here.

In the context of museum item descriptions, Poesio et al. (1999) explore the decision of what *type* of referring expression NP to use to refer to a given discourse entity, using a statistical model to choose between using a proper name, a definite description, or a pronoun. More recently, Stoia et al. (2006) attempt a similar task, but this time in an interactive navigational domain; as well as determining what type of referring expression to use, they also try to learn whether a modifier should be included. Cheng et al. (2001) try to learn rules for the incorporation of non-referring modifiers into noun phrases.

A number of the contributions to the 2008 GREC and TUNA evaluation tasks (Gatt et al., 2008) have made use of machine learning techniques. The GREC task is primarily concerned with the choice of form of reference (i.e. whether a proper name, a descriptive NP or a pronoun should be used), and so is less relevant to the focus of the present paper. Much of the work on the TUNA Task is relevant, however, since this also is concerned with determining the content of referring expressions in terms of the attributes used to build a distinguishing description. In particular, Fabbrizio et al. (2008) explore the impact of individual style and priming on attribute selection for referring expression generation, and Bohnet (2008) uses a nearest-neighbour learning technique to acquire an individual referring expression generation model for each person.

Other related approaches to attribute selection in the context of the TUNA task are explored in (Gervás et al., 2008; de Lucena and Paraboni, 2008; Kelleher and Mac Namee, 2008; King, 2008).

6 Conclusions

We know that people's referential behaviour varies significantly. Despite this apparent variation, we have demonstrated above that there does appear to be a reasonable correlation between characteristics of the scene and the incorporation of particular attributes in a referring expression. One way to conceptualise this is that the decision as to whether or

not to incorporate a given feature such as colour or size may vary from speaker to speaker; this is evidenced by the data. We might think of these as individual *reference strategies*; a good example of such a strategy, widely attested across many experiments, is the decision to include colour in a referring expression independent of its discriminatory power, perhaps because it is an easily perceivable and often-useful attribute. The overall approach to reference that is demonstrated by a given speaker then consists of the gathering together of a number of strategies; the particular combinations may vary from speaker to speaker, but as is demonstrated by the analysis in this paper, some of the strategies are widely used.

In current work, we are gathering a much larger data set using more complex stimuli. This will allow the further development and testing of the basic ideas proposed in this paper as well as their integration into a full referring expression generation algorithm.

References

Anja Belz and Albert Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of UCNLG+MT: Language Generation and Machine Translation*, pages 75–83, Copenhagen, Denmark.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2008. The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 183–191, Salt Fork OH, USA.

Bernd Bohnet. 2008. The fingerprint of human referring expressions and their surface realization with graph transducers. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 207–210, Salt Fork OH, USA.

Jean C. Carletta. 1992. *Risk-taking and Recovery in Task-Oriented Dialogue*. Ph.D. thesis, University of Edinburgh.

Hua Cheng, Massimo Poesio, Renate Henschel, and Chris Mellish. 2001. Corpus-based NP modifier generation. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh PA, USA.

Diego Jesus de Lucena and Ivandré Paraboni. 2008. USP-EACH: Frequency-based greedy attribute selection for referring expressions generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 219–220, Salt Fork OH, USA.

Giuseppe Di Fabbrizio, Amanda J. Stent, and Srinivas Bangalore. 2008. Referring expression generation using speaker-based attribute selection and trainable realization (ATTR). In *Proceedings of the Fifth International Natural Language Generation Conference*, Salt Fork OH, USA.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 198–206, Salt Fork OH, USA.

Pablo Gervás, Raquel Hervás, and Carlos León. 2008. NIL-UCM: Most-frequent-value-first attribute selection and best-scoring-choice realization. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 215–218, Salt Fork OH, USA.

John D. Kelleher and Brian Mac Namee. 2008. Referring expression generation challenge 2008: DIT system descriptions. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 221–223, Salt Fork OH, USA.

Josh King. 2008. OSU-GP: Attribute selection using genetic programming. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 225–226, Salt Fork OH, USA.

Massimo Poesio, Renate Henschel, Janet Hitzeman, and Rodger Kibble. 1999. Statistical NP generation: A first report. In *Proceedings of the ESSLLI Workshop on NP Generation*, Utrecht, The Netherlands.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 81–88, Sydney, Australia.

Jette Viethen and Robert Dale. 2008a. Generating referring expressions: What makes a difference? In *Australasian Language Technology Association Workshop 2008*, pages 160–168, Hobart, Australia.

Jette Viethen and Robert Dale. 2008b. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 59–67, Salt Fork OH, USA.

Jette Viethen, Robert Dale, Emiel Krahmer, Mariët Theune, and Pascal Touset. 2008. Controlling redundancy in referring expressions. In *Proceedings of the 6th Language Resources and Evaluation Conference*, Marrakech, Morocco.

Ian H. Witten and Frank Eibe. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.