Speaker-Dependent Variation in Content Selection for Referring Expression Generation

Jette Viethen
Centre for Language Technology
Macquarie University
Sydney, Australia
jette.viethen@mq.edu.au

Robert Dale
Centre for Language Technology
Macquarie University
Sydney, Australia
robert.dale@mq.du.au

Abstract

In this paper we describe machine learning experiments that aim to characterise the content selection process for distinguishing descriptions. Our experiments are based on two large corpora of human-produced descriptions of objects in relatively small visual scenes; the referring expressions are annotated with their semantic content. The visual context of reference is widely considered to be a primary determinant of content in referring expression generation, so we explore whether a model can be trained to predict the collection of descriptive attributes that should be used in a given situation. Our experiments demonstrate that speaker-specific preferences play a much more important role than existing approaches to referring expression generation acknowledge.

1 Introduction

Since at least the late 1980s, referring expression generation (REG) has been a key topic of interest in the natural language generation community (see, for example, (Dale, 1989; Dale and Haddock, 1991; Dale and Reiter, 1995; van der Sluis, 2001; Krahmer and Theune, 2002; Krahmer et al., 2003; Jordan and Walker, 2005; van Deemter, 2006; Gatt and van Deemter, 2006; Kelleher and Kruijff, 2006)); and it has recently served as the focus for the first major evaluation efforts in natural language generation (see, for example, (Belz et al., 2009; Gatt et al., 2009)). 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 accepted as involving 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*.

Most existing REG algorithms rely on handcrafted decision procedures whose behaviour is either entirely deterministic (Dale, 1989; Dale and Haddock, 1991; Gardent, 2002) or can be influenced to some degree using parameters such as preference orderings or cost functions over the available properties in order to choose those that should appear in a referring expression (Dale and Reiter, 1995; van der Sluis, 2001; Krahmer and Theune, 2002; Krahmer et al., 2003; van Deemter, 2006; Gatt and van Deemter, 2006; Kelleher and Kruijff, 2006). However, only very limited attempts have been made to determine how these parameters should best be instantiated in order to allow an algorithm to mimic human-produced referring expressions. Furthermore, the results of recent evaluation exercises (Gupta and Stent, 2005; Viethen and Dale, 2006; Belz and Gatt, 2007; Gatt et al., 2007; Gatt et al., 2008) show that none of these algorithms can be considered an accurate model of human production of referring expressions in any of their instantiations.

In this paper, we take a speaker-oriented perspective on REG that is aimed in part at exploring the factors that impact on the choices that humans make when they refer, and ultimately at finding models for REG which can claim at least a certain level of cognitive plausibility by being able to replicate human referring behaviour. To this end we use two large corpora of referring expressions to train machine learning models on the task of content determination. The larger of these corpora is being introduced for the first time here. We first attempt to build models that are able to predict the content of a referring expression based only on the visual characteristics of the surrounding scene. We then contrast the results of this experiment to those of a second set of experiments in which the machine learner was told which participant had

Jette Viethen and Robert Dale. 2010. Speaker-Dependent Variation in Content Selection for Referring Expression Generation. In *Proceedings of Australasian Language Technology Association Workshop*, pages 81–89

produced each description. Our results show that, while there is too much variation in the data to reliably predict the content of a referring expression based on the visual features of a scene alone, much of this variation can be accounted for by additionally taking into account participant-specific preferences. Even models based on the identity of the participants alone, while not as successful as the models based solely on scene characteristics, performed surprisingly well, underlining the importance of speaker preferences in the choice of semantic content for referring expressions.

In Section 2, we provide an overview of previous work relevant to the approach we take in this paper. Section 3 describes the two corpora that we use for training and testing our models. Section 4 details the experimental setup we used, and in Section 5 we discuss the results of our experiments. Finally, in Section 6, we summarise the key conclusions of this work and point to some future research directions we aim to pursue.

2 Related Work

There exist a number of approaches to the use of machine learning in referring expression generation, although they are typically focussed on aspects of the problem that are distinct from those considered here.

Poesio et al. (1999) addressed the decision of what *type* of NP to use to refer to a given discourse entity in the contexts of museum item descriptions and pharmaceutical information leaflets. They used a statistical model to choose between a large set of NP types, including proper names, definite descriptions, or pronouns. More recently, Stoia et al. (2006) attempted a similar task, but in an interactive navigational domain; as well as deciding what type of referring expression to use, they trained decision trees to determine whether a modifier should be included. Cheng et al. (2001) tried to learn rules for the incorporation of non-referring modifiers into noun phrases. In a domain of spoken negotiations over apartment furniture, Jordan and Walker (2005) used features based on different models of discourse theory to learn rules about which attributes to include in a referring expression. The functions performed by the referring expressions in their corpus went far beyond the simple identification task at hand in our corpora, and they had to take account of a variety of discourse-related factors impacting on their data.

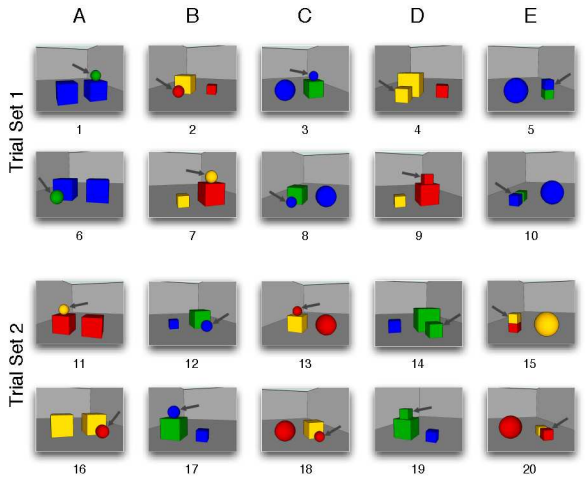

Figure 1: The 20 stimulus scenes for GRE3D3.

A number of the contributions to the 2008 and 2009 GREC and TUNA evaluation tasks 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 (Gatt et al., 2008) 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) explored the impact of individual style and priming on attribute selection for referring expression generation, and Bohnet (2008; 2009) used 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 Namee, 2008; King, 2008; Hervás and Gervás, 2009; de Lucena and Paraboni, 2009).

3 Two Corpora of Referring Expressions

3.1 Stimulus Design

The experiments in this paper are based on two corpora of human-produced referring expressions. The referring expressions were elicited by showing participants small visual scenes containing a number of simple abstract objects. One of the objects was marked by an arrow to indicate its status as the target referent to be described.

One of the initial intentions underlying both corpus collections was to investigate the condi-

tions under which participants use spatial relations to describe the target referent. Therefore, the design of all scenes is carefully controlled so that the use of relations is encouraged, but not strictly necessary in order to identify the referent. In particular, the target referent is always placed on top of or directly adjacent to a second object, which we call the *landmark* object. Each object is either a ball or a cube, large or small, and of one of two colours. The landmark object is always a cube in order to avoid unnatural looking situations where the target object would be balanced on top of a ball.

The main difference between the stimuli used to collect the two corpora lies in the number of objects contained in the scenes. The first corpus, GRE3D3,[1] has been described in detail elsewhere (Viethen and Dale, 2008); here we only summarise the key points. The stimulus scenes used to collect GRE3D3 are shown in Figure 1; they contain three objects each.

The stimuli used for the second corpus, GRE3D7,[2] contained seven objects. One of these objects was always placed on its own to one side of the scene; the remaining six appeared as three pairs of directly adjacent objects. The target was always a member of the most central pair, and one of the other pairs had the same spatial relation as that holding between the target and the landmark object. The 32 stimuli scenes for GRE3D7 are shown in Figure 2. Their design was balanced for four within-participant factors and one between-participant factor, which were chosen based on the assumption that they might impact on the use of spatial relations. These factors were the size of the landmark object, the commonness of the landmark's size (based on the number of objects sharing the landmark's size), the type of relation holding between the target and the landmark object (vertical or lateral), and two Boolean factors capturing whether the landmark and the target shared size and colour.

3.2 Procedure and Participants

The corpora were collected in two separate self-paced on-line language production experiments. Participants were asked to describe the target referent in each scene in a way that would enable another party looking at the same scene to pick it out

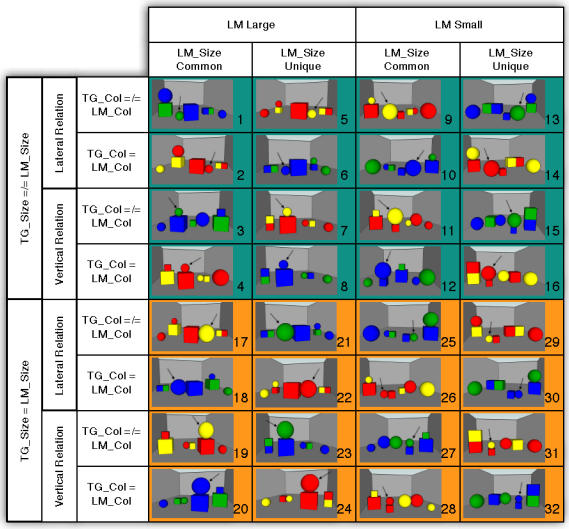

Figure 2: The 32 stimulus scenes for GRE3D7. The top half constitutes Trial Set 1 and the bottom half is Trial Set 2.

from the other objects. The scenes were presented consecutively above a text box into which the participants were required to type a description before clicking 'DONE' to move on to the next scene. In the GRE3D3 collection experiment, the scenes were presented in a preset order directly following each other. For the GRE3D7 experiment, each stimulus scene was preceded by a filler scene. The filler scenes were designed to distract the participants from noticing the similarities between the stimulus scenes. Additionally, the order in which the stimuli and the filler scenes were presented was randomised before each trial.

To encourage the use of fully distinguishing referring expressions, participants were told that they had only one chance at describing the object. After being presented with all the scenes in the trial, participants were asked to complete an exit questionnaire, which asked for their opinion on whether the task became easier over time, and any other comments they might wish to make.

The data from 63 participants in the GRE3D3 collection exercise and from 280 participants in the GRE3D7 collection exercise were used to form the final corpora. A small amount of data from both collections were discarded because the participants did not complete the whole experiment or clearly had not understood the instructions correctly. All participants were self-reported native English speakers.

Both sets of stimuli were subdivided into two trial sets and each participant saw only one of

[1] GRE3D3 stands for **G**enerating **R**eferring **E**xpressions in **3D** scenes with **3** objects.

[2] GRE3D7 stands for **G**enerating **R**eferring **E**xpressions in **3D** scenes with **7** objects.

	Content Pattern	Example Description	% Relative Frequency GRE3D3	GRE3D7
R	⟨tg_size, tg_col, tg_type⟩	the small blue ball	22.70	47.88
D	⟨tg_col, tg_type⟩	the blue ball	27.30	36.70
W	⟨tg_size, tg_col, tg_type, rel, lm_size, lm_col, lm_type⟩	the small blue ball on top of the large green cube	4.76	5.31
F	⟨tg_col, tg_type, rel, lm_col, lm_type⟩	the blue ball on top of the green cube	7.78	2.70
T	⟨tg_size, tg_col, tg_type, rel, lm_col, lm_type⟩	the small blue ball on top of the green cube	4.92	2.08
I	⟨tg_col, tg_type, rel, lm_size, lm_col, lm_type⟩	the blue ball on top of the large green cube	1.90	1.03
ZF	⟨tg_type⟩	the ball	8.25	0.07
Z	⟨tg_size, tg_type⟩	the small ball	4.44	0.38
N	⟨tg_size, tg_col, tg_loc, tg_type⟩	the small blue ball in the left	0.32	0.87
ZK	⟨tg_type, rel, lm_type⟩	the ball on top of the cube	3.49	0.40

Table 1: The ten most common content patterns that occur in both GRE3D3 and GRE3D7.

these trial sets. So, each participant in the GRE3D3 collection provided ten descriptions, while each GRE3D7 participant described 16 stimulus scenes. This resulted in 630 GRE3D3 descriptions (30 for each scene in Trial Set 1, and 33 for each scene in Trial Set 2) and 4480 GRE3D7 descriptions (140 for each stimulus scene).

3.3 Annotation of Semantic Content

In order to be able to analyse the semantic content of the referring expressions, we annotated the attributes and relations contained in each of them. The attributes that participants used in the referring expressions contained in the two corpora, and their possible values, are as follows:

- type [ball, cube]
- colour [blue, green, red, yellow]
- size [large, small]
- location [right, left, front, top]
- relation [on-top-of, in-front-of, left-of, right-of]

In our annotations, each attribute is prefixed by either tg or lm to mark whether it pertains to the target or the landmark object. For example, tg_size indicates that the size of the target was mentioned. This results in nine component properties.[3]

Each description contained in the GRE3D3 and GRE3D7 corpora can be characterised in terms of a *content pattern* defined by the presence or absence of each of these nine component properties. Table 1 lists the ten most common of these content patterns along with example descriptions and the relative frequency with which these patterns occurred in each corpus. 37 different content patterns can be found across the two corpora; the GRE3D3 corpus contains 31 of these 37 content patterns, four more than the much larger GRE3D7 corpus. 21 of the patterns occur in both corpora.

[3] As noted by one reviewer, the ethno-cultural background of speakers can have a large impact especially on the use of spatial information. The data would look very different if it had been collected from speakers of languages that mostly make absolute reference to points of the compass rather than using relative information such as 'left' and 'right'.

4 Experimental Setup

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 content patterns in this data from the contexts in which they were produced. To explore this, we define a number of features that capture the relevant aspects of the visual context in our stimulus scenes. Importantly, these features are general enough to be able to capture both GRE3D3 and GRE3D7 scenes. We use two types of features: *direct property features*, which simply record the attribute value of a certain object in the scene, and *comparative features*, which compare the attribute values of one object to those of the other objects. In a second step, we additionally include Participant_ID as a scene-independent feature. The complete list of 12 features used is shown in Table 2.

The features pay particular attention to the properties of the target and the landmark objects for two reasons: firstly, the nature of the task is such that these two objects can be expected to be closest to the participant's focus of attention; and secondly, these are the only two objects that can be identified as corresponding to each other across all scenes, in particular in the GRE3D7 stimuli.

As direct property features we use the type of spatial relation holding between target and landmark, as people generally show a preference for

	Attribute	Explanation	Values
direct property features	TG_Size	size of the target object	small, large
	LM_Size	size of the landmark object	small, large
	Relation_Type	type of relation between target and landmark	horizontal, vertical
comparative features	Num_TG_Size	number of objects of same size as the target	numeric
	Num_LM_Size	number of objects of same size as landmark	numeric
	TG_LM_Same_Size	target and landmark share size	Boolean
	Num_TG_Col	number of objects of same colour as target	numeric
	Num_LM_Col	number of objects of same colour as landmark	numeric
	TG_LM_Same_Col	target and landmark share colour	Boolean
	Num_TG_Type	number of objects of same type as target	numeric
	Num_LM_Type	number of objects of same type as landmark	numeric
	TG_LM_Same_Type	target and landmark share type	Boolean
	Participant_ID	ID number of the description giver	alphanumeric

Table 2: The features and their value formats.

vertical relations over horizontal ones (Lyons, 1977; Gapp, 1995; Bryant et al., 2000; Landau, 2003; Arts, 2004; Tenbrink, 2005), and the sizes of these two objects. We do not include colour or type as features because the actual values of these attributes are unlikely to have an impact on their use. Rather, we expect the proportion of objects sharing these properties, captured in the comparative features, to be of importance. This is different for size, as a large object stands out more from its surroundings than a small one, even independently of the sizes of the other objects. location is not included as it was almost constant across all scenes and can therefore not be used to distinguish between them.

We used the C4.5 decision tree learning algorithm (Quinlan, 1993) implemented in the Weka workbench (Witten and Frank, 2005). We tested both pruned and unpruned trees, but in what follows we comment on the results of the unpruned trees only where they are different from those of the pruned trees. Decision tree pruning is a post-training step that simplifies the trees to reduce over-fitting to the training data. This is especially relevant if the trained models are used on unseen data. However, if the ability of a feature set to characterise a set of natural data is at question, unpruned trees can also be of interest.

5 Results and Discussion

In the following, the fit of the trained models is measured by the Accuracy with which they predict held-out test data or characterise the training data. It is defined as the number of instances predicted correctly divided by the total number of instances in the test or training set.

5.1 Content Selection Based on Scene Characteristics

The Accuracy results achieved by the models trained on the scene-based feature set, without taking into account Participant_ID, are shown in Table 3. As a baseline we report the success rate of a model that simply chooses the majority class in each case. We used three different test methods: (1) testing on the complete training set shows how well the learned model characterises the data and thereby gives an indication of the extent to which the chosen features can explain the variation in the data; (2) ten-fold cross-validation is used to assess the ability of the learned model to generalise to unseen data; and finally, (3) cross-corpus testing gives insights into the difference in variation between the two data sets.

Both models significantly outperform the majority class baseline in all three test methods.[4] No difference can be found between the results for testing on the training sets and cross-corpus testing. However, three interesting observations can be made from these results:

1. Training and testing on the GRE3D7 corpus achieves better results than training and testing on the GRE3D3 corpus.

2. Both the baseline and the decision trees trained on GRE3D3 perform better on GRE3D7 than on GRE3D3 itself, while the GRE3D7-trained models achieve the lowest results when tested on GRE3D3.

3. Overall, none of the decision trees achieve very high Accuracy levels.

[4]We used χ_2 with a maximum $p<.05$ for all significance tests in this paper.

training corpus	test method	maj. class baseline	pruned tree
GRE3D3	training set	27.30%	**46.51%**
	10 fold X	27.30%	**46.51%**
	cross-corpus	36.70%	**47.88%**
GRE3D7	training set	47.88%	**64.93%**
	10 fold X	47.88%	**64.71%**
	cross-corpus	22.70%	**36.98%**

Table 3: Accuracy for the models purely based on scene characteristics. (Bold values are statistically significantly different from the baseline.)

The first two of these points indicate that the content of the referring expressions found in the GRE3D7 corpus is easier to predict than that in the GRE3D3 corpus, a fact that was already foreshadowed by the lower number of different content patterns contained in GRE3D7. The second point in particular shows that the predicted usage patterns for the different content patterns for GRE3D7 are subsumed by the GRE3D3 usage patterns.

One might consider these results to be slightly surprising, as the GRE3D7 corpus with its much larger participant base and size could have been expected to contain more variation than GRE3D3. It is possible that the filler items used in GRE3D7 prevented the participants from noticing how similar their responses to the stimulus scenes were, and thereby also prevented them from intentionally varying the content in their descriptions. A second, related, factor could be that the slightly more complex GRE3D7 scenes forced participants to concentrate on the task more, which also would lead to a reduced number of intentionally-varied descriptions.

From the third observation above we conclude that neither of the learned decision trees are able to accurately model the referring behaviour displayed by the participants in our corpora. In fact, both models predict the use of only two content patterns, patterns R and D, the two most common ones in both data sets, as shown in Table 1. The tree trained on GRE3D3 is shown in Figure 3: it only has three nodes. The GRE3D7-trained tree is at 15 nodes more complex, but nonetheless only predicts the same two most common patterns.

The overall low performance of the models might either be due to some of the variation in the data being in fact unpredictable (due to factors that we did not capture in the collection experiments) or random, or it may indicate that the features we made available to the machine learning algorithm were not sufficient to model the variation.

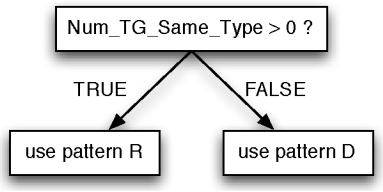

Figure 3: The participant-insensitive decision tree trained on GRE3D3.

5.2 Participant-Dependent Modelling

Based on observations we made in (Viethen and Dale, 2006) for a different data set, we hypothesise that one main factor that might be at play in producing the variation in the two corpora used here are the differing preferences of the individual participants. We therefore introduced the feature Participant_ID and carried out two further experiments: first, we tested the predictive ability of this feature on its own by removing all other features from the set provided to the machine learner; and second, we combined the scene-based features with the Participant_ID feature, in order to assess the extent to which the individual participants were taking the features of the scene into account when referring to the target referents.

Table 4 compares the results of the second two experiments to those of the participant-insensitive decisions trees from the previous section.[5]

We firstly observe that the size of the learned decision trees, measured in terms of the number of nodes they contain, increases dramatically when Participant_ID is taken into account, even when the other, scene-based, features are also available. This indicates that, when given the option to use this feature, the machine learner chooses to do so in every case, demonstrating the usefulness of Participant_ID in characterising our data.

The trees based on Participant_ID alone also achieved surprisingly good performance, although these trees are forced to choose one content pattern for all descriptions produced by a given participant. Only the Accuracy of the tree trained on GRE3D3 was significantly lower than that of the corresponding participant-insensitive tree; the other scores are surprisingly close to those based

[5]Because the participants in the two data collection exercises were not the same, cross-corpus testing of the participant-sensitive models is not possible.

training corpus	test method	+[scene features] −Participant_ID pruned Acc	pruned nodes	−[scene features] +Participant_ID n/a Acc	n/a nodes	+[scene features] +Participant_ID pruned Acc	pruned nodes	+Participant_ID unpruned Acc	unpruned nodes
GRE3D3	training set	46.51%	3	41.91%	64	**91.27%**	415	**98.10%**	573
	10 fold X	46.51%		**31.11%**		54.44%		**57.61%**	
GRE3D7	training set	64.93%	15	62.28%	281	**82.59%**	1023	**93.77%**	2798
	10 fold X	64.71%		57.12%		67.01%		63.71%	

Table 4: Accuracy and tree size for the models based on scene and participant information. (Bold values are statistically significantly different to the participant-insensitive trees.)

on scene features only.[6]

Combining the scene-based features with Participant_ID gives better results than either of the two exclusive models achieve. To the best of our knowledge, their cross-validation scores are also higher than any Accuracy scores reported in the literature for any existing algorithm instantiated with a set parameter setting.[7] However, in 10-fold cross-validation, only the unpruned GRE3D3 model achieves a statistically significant improvement over the participant-insensitive model. When testing on the training set, the pruned and unpruned trees for both corpora vastly outperform the models that do not take participant preferences into account. In particular, the Accuracy scores achieved by the unpruned models are very high.

These results confirm the hypothesis that speaker preferences play a very important role in shaping the semantic content of referring expressions in identification tasks. Trees using Participant_ID as the only feature perform surprisingly well, and the trees that take account of both the features of the scene and the preferences displayed by individual speakers are able to characterise our two data sets with very high accuracy. Our particular choice of scene-based features is also supported by these results, as they do seem to capture the factors that individual speakers rely on when they build referring expressions.

The fact that they only achieve high scores if tested directly on the training set shows that these models are very specific to the data they were trained on, and would not necessarily generalise well to unseen data. A likely explanation for the large differences between the cross-validation results and results on the training set is the low number of instances per participant in both corpora. We have ten descriptions from each participant in the GRE3D3 corpus and 16 in GRE3D7, and neither of the corpora contains multiple descriptions from the same participant for a given stimulus.

[6]Note that pruning has no effect on trees using only one feature, in this case Participant_ID.

[7]This comparison must be viewed with caution, as the other evaluations were carried out on different test corpora.

6 Conclusions and Future Work

This paper is based on the view that a primary consideration in the study of REG should be the development of systems that are able to explain and replicate the semantic content found in human data. We hold this view for two reasons: firstly, such systems can aid the exploration of factors that impact on the semantic choices that people make when they refer and ultimately might be able to claim some level of psychological reality; and secondly, generating the same referring expressions as humans can also serve a utilitarian purpose, as only human-like reference is likely to be accepted as fully natural by listeners.

We have chosen a straightforward approach to building REG models that take into account what people do by training decision trees on two human-produced corpora of distinguishing descriptions in visual scenes. We defined a set of features to capture the relevant visual aspects of the stimuli used in the data collection exercises for the two corpora. In our first experiment we established that decision trees trained using these features are able to outperform a majority class baseline, but are not able to replicate a large enough proportion of the data to be considered accurate models of human reference behaviour. In a second experiment we added the Participant_ID feature, which allowed the machine learner to establish participant-specific behaviour patterns. Trees based on this feature alone achieved surprisingly good results, and the participant-sensitive trees which also took into account the features of the scene achieved much higher Accuracy scores than

the participant-insensitive trees.

The main conclusion we draw from these experiments is that speaker-dependent variation is one of the most important factors shaping content selection processes in the referring behaviour of humans. This is an observation that has been overlooked in the development of most existing algorithms for REG. However, if our aim is to build algorithms that are able to accurately model corpora of human referring expressions, as was the case in the recent public evaluation campaigns in REG (Belz and Gatt, 2007; Gatt et al., 2008; Gatt et al., 2009), then we cannot ignore this fact.

Our next step is to take this work further by training individual models for each speaker. Such speaker-specific trees will allow us to explore the different strategies that people follow when they refer, and to compare the strategies of different speakers to each other. We think it unlikely that every individual speaker is idiosyncratic in this regard; our hypothesis is that it will be possible to use automatic clustering techniques to identify groups of people who follow the same strategies. Such clusters can then be used to make predictions that are sensitive to between-participant differences while benefitting from the commonalities in people's behaviour. It might also be interesting to see if non-linguistic characteristics of speakers, such as age, gender, and social or cultural background, can account for some of the between-participant variation in reference behaviour.

In a second strand of work we are exploring an alternative approach to learning human reference behaviour from this data. We are training attribute-specific trees that make binary decisions about the use of each individual attribute in a given reference situation, instead of predicting whole content patterns. The attribute-specific trees for a given participant can then be combined into a *speaker profile* predicting complete referring expressions produced by this speaker.

References

Anja Arts. 2004. *Overspecification in Intructive Texts*. Ph.D. thesis, University of Tilburg, The Netherlands.

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*, 75–83, Copenhagen, Denmark.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. The GREC Main Subject Reference Generation Challenge 2009: Overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, 79–87, Singapore.

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*, 207–210, Salt Fork OH, USA.

Bernd Bohnet. 2009. Generation of referring expression with an individual imprint. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 185–186, Athens, Greece.

David J. Bryant, Barbara Tversky, and M. Lanca. 2000. Retrieving spatial relations from observation and memory. In E. van der Zee and U. Nikanne (Eds.), *Cognitive interfaces: Constraints on linking cognitive information*, 94–115. Oxford University Press, Oxford, UK.

Hua Cheng, Massimo Poesio, Renate Henschel, and Chris Mellish. 2001. Corpus-based NP modifier generation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh PA, USA.

Robert Dale and Nicolas Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 68–75, Vancouver BC, Canada.

Diego Jesus de Lucena and Ivandré Paraboni. 2008. USP-EACH: Frequency-based greedy attribute selection for referring expressions generation. In *Proceedings of the 5th International Natural Language Generation Conference*, 219–220, Salt Fork OH, USA.

Diego Jesus de Lucena and Ivandré Paraboni. 2009. USP-EACH: Improved frequency-based greedy attribute selection. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 189–190, Athens, Greece.

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 5th International Natural Language Generation Conference*, 211–214, Salt Fork OH, USA.

Klaus-Peter Gapp. 1995. Angle, distance, shape, and their relationship to projective relations. In *Proceedings of the 17th Conference of the Cognitive Science Society*, 112–117, Pittsburgh PA, USA.

Claire Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 96–103, Philadelphia PA, USA.

Albert Gatt and Kees van Deemter. 2006. Conceptual coherence in the generation of referring expressions. In *Proceedings of the 21st COLING and the 44th ACL Conference*, 255–262, Sydney, Australia.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation*, 49–56, Schloß Dagstuhl, Germany.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference*, 198–206, Salt Fork OH, USA.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 174–182, Athens, Greece.

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 5th International Natural Language Generation Conference*, 215–218, Salt Fork OH, USA.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, 1–6, Brighton, UK.

Raquel Hervás and Pablo Gervás. 2009. Evolutionary and case-based approaches to REG: NIL-UCM-EvoTAP, NIL-UCM-ValuesCBR and NIL-UCM-EvoCBR. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 187–188, Athens, Greece.

Pamela W. Jordan and Marilyn A. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

John Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st COLING and the 44th ACL Conference*, 1041–1048, Sydney, Australia.

John D. Kelleher and Brian Mac Namee. 2008. Referring expression generation challenge 2008: DIT system descriptions. In *Proceedings of the 5th International Natural Language Generation Conference*, 221–224, Salt Fork OH, USA.

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

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, 223–264. CSLI Publications, Stanford CA, USA.

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

Barbara Landau. 2003. Axes and direction in spatial language and spatial cognition. In Emilie van der Zee and Jon M. Slack (Eds.), *Representing Direction in Language and Space*, 18–38. Oxford University Press, Oxford, UK.

John Lyons. 1977. *Semantics*, volume 2. Cambridge University Press.

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.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco CA, USA.

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*, 81–88, Sydney, Australia.

Thora Tenbrink. 2005. Semantics and application of spatial dimensional terms in English and German. Technical Report Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition, No. 004-03/2005, Universities of Bremen and Freiburg, Germany.

Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

Ielka van der Sluis. 2001. An empirically motivated algorithm for the generation of multimodal referring expressions. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Student Session*, 67–72, Toulouse, France.

Jette Viethen and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, 63–70, Sydney, Australia.

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

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco CA, USA.