What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions

Albert Gatt
Institute of Linguistics
University of Malta
albert.gatt@um.edu.mt

Patrizia Paggio
Institute of Linguistics
University of Malta
patrizia.paggio@um.edu.mt

Abstract

Pointing gestures are pervasive in human referring actions, and are often combined with spoken descriptions. Combining gesture and speech naturally to refer to objects is an essential task in multimodal NLG systems. However, the way gesture and speech should be combined in a referring act remains an open question. In particular, it is not clear whether, in planning a pointing gesture in conjunction with a description, an NLG system should seek to minimise the redundancy between them, e.g. by letting the pointing gesture indicate locative information, with other, non-locative properties of a referent included in the description. This question has a bearing on whether the gestural and spoken parts of referring acts are planned separately or arise from a common underlying computational mechanism. This paper investigates this question empirically, using machine-learning techniques on a new corpus of dialogues involving multimodal references to objects. Our results indicate that human pointing strategies interact with descriptive strategies. In particular, pointing gestures are strongly associated with the use of locative features in referring expressions.

1 Introduction

Referring Expression Generation (REG) is considered a core task in many NLG systems (Krahmer and van Deemter, 2012). Typically, the REG task is defined in terms of identification: a referent needs to be unambiguously identified in a discourse, enabling the reader or listener to pick it out from among its potential distractors. Most work in this area has focused on algorithms that select the content for definite descriptions (Dale, 1989; Dale and Reiter, 1995), or on the best form for a referring expression given the discourse context, for example, whether it should be a full definite description, a reduced one, or a pronoun (McCoy and Strube, 1999; Callaway and Lester, 2002; Krahmer and Theune, 2002).

Less attention has been payed to the role of gestures in referring actions and the way these can be coupled with discursive strategies for referent identification. This question becomes particularly important in the context of multimodal systems, for example, those involving embodied conversational agents, where the 'naturalness' of an interaction hinges in part on the appropriate use of embodied actions, including referring actions. Multimodal strategies can also make communication more efficient. For example, Louwerse and Bangerter (2010) found that the use of pointing gestures resulted in significantly faster resolution of ambiguous referring expressions; crucially, this result was replicated when the pointing gesture was artificially generated, rather than made by a human.

Like human communicators, embodied agents need the ability to plan multimodal referring acts, combining both linguistic reference and pointing. An important question is whether these two components of a referring act should be planned in order to minimise redundancy between them or not. For example, given that a pointing gesture can efficiently locate a target referent in a visual domain, should an accompanying description avoid mentioning locative properties, thereby minimising redundancy? This question is the main focus of this paper. However, it bears on a deeper issue, of relevance to the architecture of multimodal systems (and the cognitive architectures whose behaviours such systems seek to emulate): Should gestural and descriptive strategies be viewed as separate (implying that a REG module can plan its linguistic referring expressions more or less in-

Proceedings of the 14th European Workshop on Natural Language Generation, pages 82–91,
Sofia, Bulgaria, August 8-9 2013. ©2013 Association for Computational Linguistics

dependently of whether a pointing gesture is also used) or should they be viewed as tightly coupled? If they are indeed coupled, are there any features of a linguistic description (for example, an object's location) which are excluded when a pointing gesture is used, or are linguistic features always redundant with pointing?

The present paper addresses these questions in a data-driven fashion, using a multimodal corpus of dialogues collected specifically to study referring actions at both the linguistic and gestural levels. We focus on pointing (that is, *deictic*) gestures directed at an intended referent (as opposed to, say, iconic gestures) and investigate the extent to which pointing interacts with linguistic means for referent identification. Following an overview of previous work on pointing and reference (Section 2) and a description of the corpus (Section 3), we describe a number of machine-learning experiments that address the main empirical question (Section 4), concluding with a discussion.

2 Background: Pointing and describing

There is a growing consensus in the psycholinguistic literature, especially following the work of McNeill (McNeill, 1985), that gesture and language share a number of underlying mental processes and are therefore coupled to a significant degree. This view is in part based on the observation that gestures are temporally coupled with speech and contribute meaningfully to the achievement of a communicative intention (McNeill and Duncan, 2000). For instance, in the example below, extracted from our corpus (see Section 3), a speaker identifies a landmark (composed of a collection of five circles) on a map through a combination of a pointing gesture and the mention of the size and colour of the elements making up the landmark.

(1) there's a group of five large red ones [points]

In this case, the pointing gesture further contributes to the communicative aim of identifying the cluster of five objects, in tandem with the visual features mentioned in the description. McNeill's proposal (McNeill and Duncan, 2000) is that speech and gesture should be considered as the joint outcome of the language production process, rather than as outcomes of separate processes. Various models have been proposed which are more or less congruent with this view. For example, de Ruiter (2000) proposes that the two modalities are planned together at early stages of conceptualisation during speech production, while Kita and Özyürek (2003) suggest that gestures are planned by spatio-motoric processes which differ from the planning of speech production, but interact with it at particular points.

Recent computational work has also taken these ideas on board. For example, Kopp et. al. (2008) describe a system for the concurrent planning and generation of gesture and speech, whose architecture is inspired by Kita and Özyürek (2003) and which makes use of 'multimodal concepts' (inspired by McNeill's 'growth points') combining both propositional and visuo-spatial properties. This contrasts with earlier architectures, such as that proposed by André and Rist (1996), where generation of text and gesture is undertaken by separate modules communicating with a central planner.

The idea that the planning of language is tightly coupled with that of gesture raises the possibility that the two modalities may overlap to different degrees. Gesture may be completely redundant with speech, or may encode aspects of the communicative intention that are not included in the linguistic message itself. This raises an interesting question for multimodal REG: are there features of objects that tend to be mentioned in tandem with a pointing gesture; if so, which are they? For example, the reference in (1) mentions the size and colour of the landmark, but not its location, possibly suggesting that the speaker relied on pointing to convey the 'where' of the target referent, as opposed to the 'what', which is conveyed by the description. This, however, is not the case in the example below, where pointing is accompanied by a mention of the referent's location.

(2) [...] the red ones directly to the left [...]
[points]

There are at least two views on the relationship between pointing and describing (de Ruiter et al., 2012). On the one hand, the *trade-off* hypothesis holds that the decision to use a pointing gesture depends on the effort or 'cost' involved (the further away from the speaker and the smaller a referent is, the more costly it would be to point at it), compared to the effort involved in describing a referent linguistically.

On the other hand, pointing and (some aspects of) describing might proceed hand in hand, so that there is some degree of redundancy between the two modalities. Under this view, pointing may be chosen not based on (low) cost assessment but as part of a specifically multimodal cognitive strategy.

Evidence for the trade-off hypothesis is reported by Bangerter (2004), who found that, as pointing became easier in a task-oriented dialogue (because the distance between the speaker and the referent was shorter), there was a decrease in verbal effort, as measured by the number of words produced, as well as a decrease in the use of locative and visual features such as colour. Piwek (2007) also found that referring acts accompanied by pointing tended to include descriptions containing fewer properties than those which were not. These results are compatible with a view of the speaker/generator as essentially seeking to minimise effort in the communicative act, adopting the easiest available strategy that will not compromise communicative success (Beun and Cremers, 1998).

Similar results are reported by van der Sluis and Krahmer (2007), who model the trade-off hypothesis in a multimodal REG algorithm based on the graph-based framework of Krahmer et. al. (2003). The algorithm chooses to use pointing gestures, with various degrees of precision, depending on their cost relative to that of features that can be used in a linguistic description.

There is also evidence against the trade-off model. Recent experimental work by de Ruiter et. al. (2012) showed that the tendency for speakers to point was unaffected by the difficulty of referring to an object using linguistic features, although pointing did decrease with repeated reference to the same entities. Interestingly, the authors observed a correlation between the rate of pointing and the use of locative properties of objects. This would appear to favour a model in which the linguistically describable features of objects are differentiated: speakers may be using locative properties and pointing together as part of a strategy to identify the 'where' of an object. This is in line with the observation by Louwerse and Bangerter (2010) that, in visual domains, using pointing gestures with locative expressions increases the speed with which references are resolved.

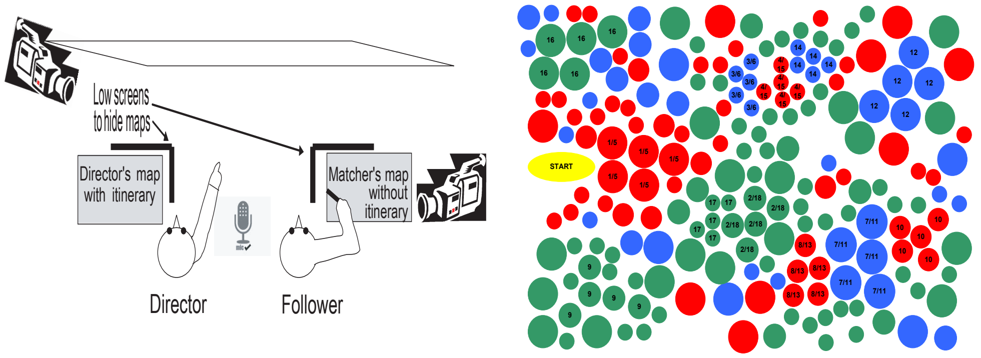
The evidence from de Ruiter et. al would seem to contradict the assumptions underlying current multimodal REG models. As we have seen, van der Sluis and Krahmer (van der Sluis and Krahmer, 2007) assume a trade-off between speech and gesture. A similar assumption is made by Kranstedt and Wachsmuth (2005), who view pointing gesturs as mainly concerned with the 'where' of an object. Their algorithm, which underlies the planning of multimodal references by a virtual agent, extends the Incremental Algorithm (Dale and Reiter, 1995) as follows. Given an object in a 3D space, the algorithm first considers the possibility of producing an unambiguous pointing gesture; failing this, a pointing gesture covering the intended referent and some of its surrounding distractors may be planned. In the latter case, the algorithm then integrates other features of the object (e.g. its colour), in an effort to exclude the distractors that remain within the scope of the ambiguous point. One of the claims underlying this model is that 'absolute' location, which is covered by pointing, is given first preference after pointing itself, with other features of a referent being considered afterwards, in a preference order that will only use relative location if all other options (such as colour) are exhausted.

In summary, the empirical evidence for the relationship between pointing and describing is mixed. While the view that the planning of language in different modalities should be tightly coupled has proven useful and productive, the precise way in which the two interact in a referring act is still an open question, especially where the relationship between location and the other features of a target referent is concerned. In the remainder of this paper, we report on an empirical study that used machine learning methods with a view to establishing the relationship between descriptive features and pointing in multimodal references. Our study is not committed to a specific architecture for multimodal reference planning; rather, our aim is to establish whether pointing and describing can partly overlap in the information that they convey about a referent. Specifically, we are interested in whether the use of a description that includes spatial or locative information excludes a pointing gesture.

3 Corpus and data

The data used in this study comes from the MREDI (Multimodal REference in DIalogue) corpus (van der Sluis et al., 2008)[2], a new collection of dia-

[2] We intend to make this corpus publicly available in the near future.


Low screens to hide maps
Director's map with itinerary
Matcher's map without itinerary
Director
Follower
START

(a) Experiment setup

(b) Group circles map (numbers indicate the order in which landmarks are visited along the itinerary)

Figure 1: MREDI dialogue setup

	Feature	Name	Definition	Example
Visual	S	Size	mention of the target size	*the group of small circles*
	Sh	Shape	mention of the target shape	*the circles at the bottom*
	C	Colour	mention of the target colour	*The blue square near the red square*
Deictic/anaphoric	I	Identity	Statement of identity between the current and a previous or later target	*the red square, the same one we saw at number 5*
	D	Deixis	Use of a deictic reference	*those squares*
Locative	RP	Relative position	Position of the target landmark relative to another object on the map	*the blue square just below the red square*
	AP	Absolute position	Target position based on absolute frame of reference	*The blue circle down at the bottom*
	FP	Path references	References to non-targets on the path leading to the target.	*go east to the first tiny square, past the blue one*
	DIR	Directions	Direction-giving.	*take a right, go across and straight down*
Action	GZ Point	Gaze Pointing	Gaze at the shared map (boolean). Use of a pointing gesture (boolean).[1]	

Table 1: Features annotated in the dialogues. All features have frequency values, except for the Action features, which are boolean.

logues elicited using a task similar to the Map-Task (Anderson et al., 1991), in which a director and a follower talked about a map displayed on a wall in front of them, approximately 1 metre away. Each also had a private copy of the map; the director's map had an itinerary on it, and her task was to communicate the itinerary to the follower, who marked it on his own private map. Participants were free to interact using speech and gesture, without touching the shared map or standing up. They could see each other, but could not see each other's private maps. Figure 1(a) displays the basic experimental setup.

The maps consisted of shapes (squares or circles), with a sequence of landmarks constituting the itinerary (initially known only to the director). The maps were designed to manipulate a number of independent variables, in a balanced design:

- **Cardinality** The target destinations in the itineraries were either individual landmarks (in 2 of the maps) or sets of 5 landmarks with the same attributes (e.g., all green squares);
- **Visual Attributes:** Targets on the itinerary differed from their distractors – the objects in their immediate vicinity (the *focus area*) – in colour, or in size, or in both colour and size. The focus area was defined as the set of objects immediately surrounding a target;
- **Prior reference:** Some of the targets were visited twice in the itinerary;
- **Shift of domain focus:** Targets were located near to or far away from the previous target. Note that if two targets t_1 and t_2 were in the *near* condition, then t_1 is one of the distractors of t_2 and vice versa.

Each participant dyad did all four maps (singleton squares and circles; group squares and circles), in a pseudo-random order, alternating in the director/matcher role so that each was director for two of the maps. Figure 1(b) displays the director's map consisting of group circles. Note that the itinerary is marked by numbering the target landmarks. Landmarks with two numbers are visited twice (for example, the first landmark is marked 1, but is also marked 5, meaning that it is the first and the fifth landmark in the itinerary). During the experiment, the map was mounted on a wall and blown up to A0 size; this significantly reduced the impression of visual clutter.

Data was collected from 8 pairs of participants[3]. In the present study, we focus exclusively on the directors' utterances. These were transcribed and split up according to the landmark to which they corresponded. In case a landmark was described over multiple turns in the dialogue, each turn was annotated as a separate utterance. Utterances were annotated with the features displayed in Table 1. Broadly, features are divided into four types: (a) *Deictic/Anaphoric*, pertaining to the use of deictic demonstratives, and/or references to previously identified entities; (ii) *Visual*, that is, corresponding to a landmark's perceptual properties; (iii) *Locative*, involving a description of the object's location; and (iv) *Action*, pertaining to gesture and gaze. All features are frequencies per utterance, except for Action features, which are boolean.

Feature	Frequency	Mean	SD
S	510	0.23	0.48
Sh	252	0.10	0.40
C	603	0.30	0.50
I	249	0.10	0.40
D	375	0.17	0.43
RP	529	0.13	0.40
AP	293	0.13	0.40
FP	989	0.40	0.70
DIR	251	0.11	0.37
GZ	836		
Point	370		

Table 2: Descriptive statistics for features in the corpus

The corpus consists of a total of 2255 director's utterances. The frequency of each feature in the corpus, as well as the per-utterance mean and standard deviation (where relevant), are indicated in Table 2; note that, with the exception of Action features, all feature values are frequencies per utterance.

Type	No point (#)	Point (#)	Total
Group	907	201	1108
Singleton	978	169	1147
Total	1885	370	2255

Table 3: Frequency of occurrence of pointing gestures relative to different object types.

As expected, linguistic features are much more frequent than pointing gestures. In fact only 16.4% of the utterances in the corpus are accompanied by pointing gestures. Previous studies, such as that by Beun and Cremers (Beun and Cremers, 1998) report a higher incidence of pointing (48% overall). Note, however, that Beun and Cremers focussed exclusively on first mention descriptions (which numbered 145 in all), while our corpus includes subsequent mentions, as well as multiple consecutive references to the same object divided over several utterances (which are counted separately in our totals).

Table 3 shows frequency figures for the pointing gestures in the corpus relative to the type of object they refer to (group vs. singleton): in accordance with the trade-off theory, which predicts that larger objects should be easier to point at, we see a significant difference ($\chi^2(1) = 4.769$, $p = 0.028$) between the two types, with more pointing occurring with group objects (that is, in group maps).

4 Experiments

In much of the work discussed in Section 2, the generation of pointing gestures is viewed as dependent on physical characteristics of the referents, in other words on their being suitable for pointing. This is especially true of work related to the trade-off hypothesis, in which the costs of pointing gestures are calculated as a function of the referent object's size and its distance from the speaker. In the present paper, by contrast, we are interested in investigating the relation between pointing and linguistic means of referent identification. More specifically, we address the question to what degree the different linguistic expressions used by the speaker to refer to objects in

[3] A number of other dialogues were recorded, but were not included in the corpus because participants focused on their own private maps and never used pointing gestures, making it impossible to study the conditions under which such gestures are produced.

the MREDI dialogues, can be used to predict the occurrence of pointing gestures. Note that this question addresses the *correlation* between properties in a description and the occurrence of pointing, rather than the issue of *how* pointing and describing should be planned. Nevertheless, as we have emphasised in Section 2, the question of co-occurrence of the two referential strategies does have a bearing on architectural issues.

A first set of experiments were run in order to test the general trade-off hypothesis. We tested a number of classifiers on the task of classifying the binary feature *point*, given all the linguistic features in the corpus. More specifically, the attributes used for the classification were *MapConfl, DIR, RP, AP, FP, S, Sh, C, D, I, Point*. They are all explained and exemplified in Table 1 with the exception of *MapConfl*, which indicates whether a specific case in the data comes from a group or a singleton map. This feature was included because, as noted in the previous section, whether a target landmark was a singleton or a group made a difference, presumably because groups are larger and more visually salient. Note further that one of the Action features, *GZ* (gaze), is ignored in the experiments because it is an almost univocal predictor of pointing. Indeed, gazing is involved roughly every time *Point* has the value *y* (yes) (but not the other way round).

The experiments were run using the Weka (Witten and Frank, 2005) tool, which gives access to many different algorithms, and 10-fold cross-validation was used throughout. The results are shown in Table (4) in terms of Precision, Recall and F-measure for each of the classifiers.

Classifier	**P**	**R**	**F**
Baseline 1 (ZeroR)	0.699	0.836	0.761
Baseline 2 (OneR)	0.762	0.834	0.765
SMO	*0.699*	0.836	*0.761*
NaiveBayes	0.795	0.811	0.802
Logistic	0.806	0.84	0.808
J48	**0.829**	**0.85**	**0.833**

Table 4: Predicting pointing gestures given all the linguistic features in the corpus: classification results.

Two baselines were created to evaluate the results. The first one is provided by the ZeroR classifier, which always chooses the most frequent class, in this case *n* (no pointing gesture). The F-measure obtained by this method is somewhat high at 0.761, because there are relatively few pointing gestures in the data. The second baseline, which provides a slightly more interesting result against which to evaluate the other classifiers, is provided by OneR. It achieves an F-measure of 0.765 by predicting a pointing gesture if DIR >= 2.5, in other words if there are at least 2.5 occurrences of direction expressions in the utterance. Using this rule has the effect of predicting a few of the pointing gestures, with an F-measure on the *y* class (occurrence of pointing gestures) of 0.031.

The other four sets of results were obtained by running four different classification algorithms with the same set of attributes. Apart from SMO (an algorithm using support vector machines), all the classifiers perform better than the baseline. The best results are produced by the decision tree classifier J48, which obtains an overall F-measure of 0.833, and an F-measure of 0.421 on the *y* class. The confusion matrix generated by J48 on this data-set is shown in Table (5)

a	**b**	← classified as
1794	91	a = n
247	123	b = y

Table 5: Predicting pointing given all the linguistic features in the corpus: confusion matrix.

The model created by the decision tree classifier (J48) is quite complex (size=57 and no. of leaves=29). The first branching, which corresponds to no *AP* (Absolute Position) and no *C* (Colour), assigns *n* to as many as 1571 instances (with 115 errors). The tree is shown in Figure (2). The tree also shows that certain combinations of features are more likely to be associated with pointing gestures. These are predominantly combinations including occurrences of *AP*, or, in the absence of absolute position, combinations including positive values for *FP* (Frequency of reference on Path) and *DIR* (Direction).

The maximum entropy model, built by the logistic regression algorithm (Logistic), shows similar tendencies in that the attributes that are assigned the highest weights are *AP*, *C* and *DIR*.

These results confirm the general hypothesis that there is a strong relationship between linguistic features used in a description and pointing gestures. Indeed, it is possible to predict pointing gestures on the basis of the linguistic features used.

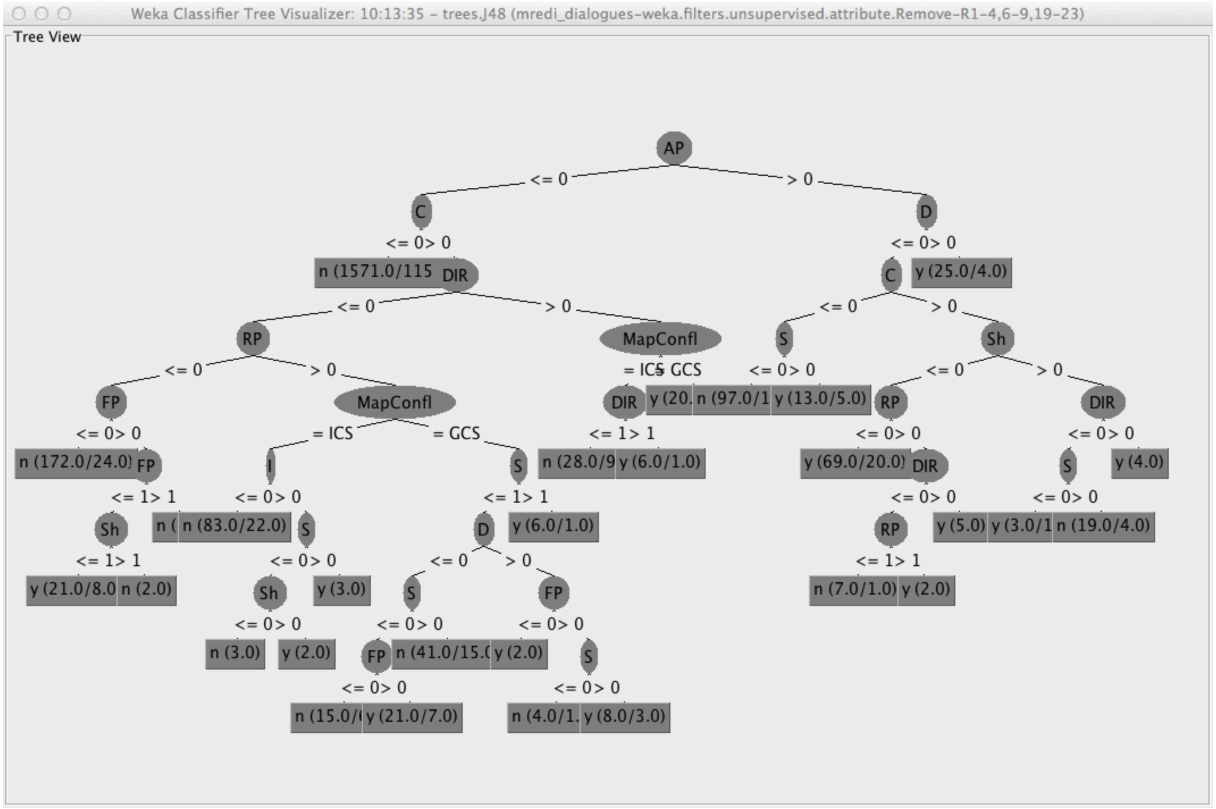


Figure 2: J48 decision tree

Classifier	P	R	F	Features
Exp1: J48	0.829	0.85	0.833	All features
Exp3: Logistic	0.806	0.84	0.808	Loc+D+I
Exp2: J48	0.835	0.851	0.806	MapConfl+Loc+D+I
Exp6: NaiveBayes	0.793	0.825	0.802	Loc
Exp4: NaiveBayes	0.764	0.804	0.779	MapConfl+Visual+D+I
Exp5: J48	0.761	0.808	0.777	MapConfl+Visual
Exp8: NaiveBayes	0.761	0.808	0.777	Visual
Exp9: NaiveBayes	0.761	0.801	0.775	Visual+D+I
Baseline 2: OneR	0.762	0.834	0.765	Dir
Exp7: F48	0.699	0.836	0.761	MapConfl+D+I
Baseline 1: ZeroR	0.699	0.836	0.761	Most freq class

Table 6: Predicting pointing gestures with different feature combinations: classification results.

In particular, the results suggest a difference between features that express locative properties and those having to do with the visual description of the same object (its colour, size and shape). More specifically, it would seem that locative features are more useful to the classifiers than visual properties.

To test this second hypothesis, we ran a series of experiments where the task was still to predict pointing gestures, but different subsets of the linguistic features were tested one at the time. For each feature combination, we run the classification using J48, Naive Bayes and the Logistic regression algorithm. In Table (6), we show the best result obtained for each feature combination. The classifiers are ordered from the most accurate to the least accurate, and the combination of features used by each of them is listed in the last column. The best results and the two baselines from the previous set of experiments are included for the sake of comparison. Note that the term *Loc* is used to refer to all the locative attributes *AP, DIR, RP, AP* and *FP*, while *Visual* refers to *S, Sh* and *C*.

The best results are those obtained when the complete feature set is used in the training. However, the next best results are achieved by the classifiers using the locative features, either alone or together with features concerning the map type, identity with a previously mentioned object and deictic reference, with an F-measure in the range 0.802–0.808. If visual features are used instead, the F-measure is in the range 0.775–0.779. The worst results are obtained if neither location nor visual description are used. Thus, although the differences between the best and the worst classifiers are not dramatic, in this data we see a tendency for the locative features to be slightly better predictors of pointing gestures than features corresponding to visual descriptions.

5 Discussion and conclusions

The automatic classification experiments described above show that to a certain extent, the pointing gestures occurring in the MREDI corpus can be predicted based on the linguistic expressions used by the speaker in conjunction with pointing. More precisely, linguistic descriptions can be used to predict about one third of the pointing gestures that speakers have produced in the corpus. This is an interesting and novel result, which not only supports the general notion that gestures and speech should be seen as tightly coupled, but also suggests that this coupling does not result in a minimisation of redundancy between the two modalities. Rather, it appears that a number of pointing gestures accompanied descriptions containing locative properties, something that contradicts the predictions of models based on the trade-off hypothesis (Kranstedt and Wachsmuth, 2005; van der Sluis and Krahmer, 2007).

There are a number of limitations of the present study, which we plan to address in future work. First, pointing gestures in our corpus were relatively scarce (16.4% of utterances were accompanied by pointing). This in part explains the relative accuracy of our baselines: predicting the majority class (that is, no pointing) in every case will clearly yield reasonable results given that the size of the class is so large. On the other hand, the relative scarcity of pointing may also indicate that pointing is somewhat more costly than linguistic description, in cognitive and physical terms. In fact, the difference we see in the number of pointing gestures between singleton and group maps also seems to confirm this assumption: in the group maps, where objects are larger, and thus more easily pointed at according to the trade-off model, there are in fact significantly more pointing gestures. The incidence of pointing may also have been affected by the nature of the domains used: although the shared maps in the experiments were large and quite close to the interlocutors, the presence of objects of the same shape may have added to the general visual clutter of the maps, making pointing less likely.

Another aspect of the data that we have not investigated is the presence of individual strategies. We know that speakers differ a lot in their use of gesturing as regards e.g. frequency, type of gesture and representation techniques. Recent models of gesture production for embodied agents are taking such differences into account (Neff et al., 2008; Bergmann and Kopp, 2009). Similarly, some speakers might have a greater preference for pointing than others. For example, Beun and Cremers (1998) note that certain speakers in their corpus explicitly stated that they had attempted to perform the task in their dialogues without pointing, in spite of their having been told that they could point. Recent data-driven experiments on referential descriptions by Dale and Viethen (Dale and Viethen, 2010), In a domain quite similar to the one used here, suggest that speakers do indeed cluster according to their preferred referential strategy. Similar assumptions have informed REG algorithms trained on the TUNA Corpus, in the context of the Generation Challenges (Gatt and Belz, 2010) (Bohnet, 2008; Di Fabbrizio et al., 2008). In future work, we plan to address this question in a multimodal context, where results by Piwek (2007) have already suggested that such individual strategies may play an important role.

The hypothesis that specific combinations of pointing and linguistic descriptions (for example, an object's colour or size) can be excluded, is clearly not borne out by the data. There is, however, a tendency for locative features to act as stronger predictors of pointing gestures. Although the trend is not very strong, it is an interesting one since it confirms the experimental results by de Ruiter et. al. reviewed earlier (de Ruiter et al., 2012). This may suggest that a pointing gesture may ultimately be planned within the same system as locative features (i.e. the decision of whether or not to point is not dependent on the decision of whether or not to describe inherent, visual properties of the object, but on whether the object's location is to be indicated). Another feature that is worth exploring further is deixis, specifically the difference between proximal and distal deictic expressions and their interaction with pointing gestures. For example, Piwek et al. (2007) found that proximal deictic expressions tend to be associated with a more intensive attentional focusing mechanism, while Bangerter (2004) also observes an association between pointing and the use of deictic expressions.

From an NLG perspective, our results suggest that decisions to generate a pointing gesture and those to select visual attributes might take place independently (perhaps in parallel, perhaps in different modules). From a cognitive perspective, it suggests two types of interaction between attention/vision and language/gesture, related to the description of the 'what' of an object and its 'where' (Landau and Jackendoff, 1993).

Finally, our study focused on the relationship between the two modalities involved in a referential act, addressing the question of redundancy between them. We have not addressed the impact of the visual properties of a target referent in relation to its surrounding objects, on the choices speakers make in these two modalities. This is a priority for future work, given that the corpus was designed to balance the presence or absence of various visual properties of an object (see Section 3). Taking this even further, it remains to be investigated, for example, whether there would be interesting differences in the relationship betwene pointing and describing between 2D scenes of the kind used here, and 3D environments of the sort used by Kranstedt and Wachsmuth (2005). Another priority is to take into account the interactive nature of the dialogues, with particular focus on the follower's feedback to the director, as an indicator of the success of referential expressions. This is another aspect of the dialogue situation that may have an impact on planning multimodal referential acts.

Acknowledgements

Special thanks are due to Ielka van der Sluis, Adrian Bangerter and Paul Piwek, who were involved in every step of the design, collection and annotation of the MREDI corpus, and who also commented on preliminary drafts of this paper.

References

A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34:351–366.

E. André and T. Rist. 1996. Coping with temporal constraints in multimedia presentation planning. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI'96).*

A. Bangerter. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419.

K. Bergmann and S. Kopp. 2009. GNetIc - using bayesian decision networks for iconic gesture generation. In A. Nijholt and H. Vilhjálmsson, editors, *Proceedings of the 9th International Conference on Intelligent Virtual Agents (LNAI 5773)*, pages 76–89. Springer.

R.J. Beun and A. Cremers. 1998. Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6(1-2):121–152.

B. 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 (INLG'08).*

C. Callaway and J. C. Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02).*

R. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

R. Dale and J. Viethen. 2010. Attribute-centric referring expression generation. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *LNAI*. Springer, Berlin and Heidelberg.

R. Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the Association for Computational Linguistics (ACL'89)*, pages 68–75.

J.P. de Ruiter, A. Bangerter, and P. Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4:232–248.

J.P. de Ruiter. 2000. The production of gesture and speech. In D. McNeill, editor, *Language and Gesture*, pages 284–311. Cambridge University Press.

G. Di Fabbrizio, A. J. Stent, and S. Bangalore. 2008. Trainable speaker-based referring expression generation. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL'08)*, pages 151–158.

A. Gatt and A. Belz. 2010. Introducing shared task evaluation to nlg: The TUNA shared task evaluation challenges. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*. Springer.

S. Kita and A. Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32.

S. Kopp, K. Bergmann, and I. Wachsmuth. 2008. Multimodal communication from multimodal thinking: Towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, 2(1):115–136.

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

E. Krahmer and K. van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

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

A. Kranstedt and I. Wachsmuth. 2005. Incremental generation of multimodal deixis referring to objects. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG'05).*

B. Landau and R. Jackendoff. 1993. what and where in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–238.

M. Louwerse and A. Bangerter. 2010. Effects of ambiguous gestures and language on the time-course of reference resolution. *Cognitive Science*, 34:1517–1529.

K.F. McCoy and M. Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of the Workshop on the Relation of Discourse/Dialogue Structure and Reference*.

D. McNeill and S.D. Duncan. 2000. Growth points in thinking for speaking. In D. McNeill, editor, *Language and Gesture*, pages 141–161. Cambridge University Press.

D. McNeill. 1985. So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371.

M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. 2008. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24.

P. Piwek, R-J. Beun, and A. Cremers. 2007. proximal and distal in language and cognition: Evidence from deictic demonstratives in dutch. *Journal of Pragmatics*, 40(4):694–718.

P. Piwek. 2007. Modality choice for generation of referring acts: Pointing vs describing. In *Proceedings of the Workshop on Multimodal Output Generation (MOG'07).*, pages 129–139.

I. van der Sluis and E. Krahmer. 2007. Generating multimodal referring expressions. *Discourse Processes*, 44(3):145–174.

I. van der Sluis, P. Piwek, A. Gatt, and A. Bangerter. 2008. Towards a balanced corpus of multimodal referring expressions in dialogue. In *Proceedings of the Symposium on Multimodal Output Generation (MOG'08).*

I.H. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, second edition.