Learning when to point: A data-driven approach

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

Patrizia Paggio
Institute of Linguistics, Uni of Malta
Centre for Language Technology, Uni of Copenhagen
patrizia.paggio@um.edu.mt

Abstract

The relationship between how people describe objects and when they choose to point is complex and likely to be influenced by factors related to both perceptual and discourse context. In this paper, we explore these interactions using machine-learning on a dialogue corpus, to identify multimodal referential strategies that can be used in automatic multimodal generation. We show that the decision to use a pointing gesture depends on features of the accompanying description (especially whether it contains spatial information), and on visual properties, especially distance or separation of a referent from its previous referent.

1 Introduction

The automatic generation of multimodal referring actions is a relatively under-studied phenomenon in Natural Language Generation (NLG). While there has been extensive research on Referring Expression Generation (REG) focusing on the choice of content in expressions such as (1) below (Dale, 1989; Dale and Reiter, 1995; Krahmer and van Deemter, 2012), their multimodal counterpart – exemplified in (2) – raises questions that go beyond these choices.

(1) the group of five large red circles

(2) there's a group of five large red ones [+pointing gesture with arm extended]

One important question concerns the appropriateness of a pointing gesture under different conditions. The relevant conditions here include both the physical or perceptual common ground shared by interlocutors (for example, what other objects are in the vicinity of the target referent, and therefore potentially confusable with it), the discursive common ground (for example, whether this object has been referred to before) and the content of the interlocutor's speech act, that is, what she chooses to say in addition to pointing. For example in (2), the speaker, who is engaged in a dialogue in which she needs to guide her interlocutor through a route on an abstract map (see Section 3 below), has chosen to use the cardinality of the referent (it is a group made up of five circles), its size, and its colour. Her choice of properties may be sufficient to distinguish it from all its distractors in the current context. However, unlike (1), (2) is a *composite utterance* consisting of two communicative modalities, each of which contributes to the communicative intention (Enfield, 2009).

This paper addresses the question of when a pointing gesture is appropriate as part of a composite, multimodal referring action. This is an important component of many multimodal generation systems, including those that communicate through embodied agents. We address this question in a data-driven manner, using a corpus of dialogues in which references have been annotated at both the level of speech and gesture. Our aim is to learn strategies for combinations of pointing and describing, as a function of perceptual and discursive features. We first summarise some relevant psycholinguistic and computational work (Section 2), before describing our corpus data (Section 3) and reporting on the machine-learning experiments conducted (Section 4). Section 5 concludes with some remarks on future work.

2 Pointing and reference

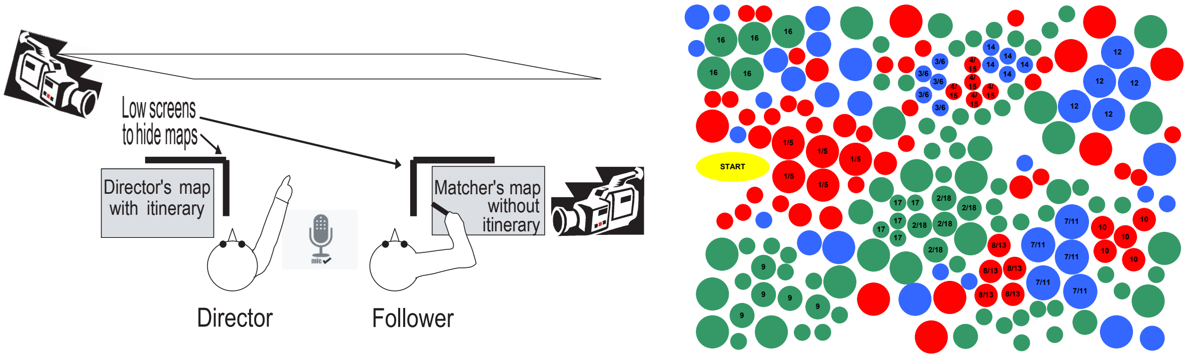
The idea that gesture and speech are planned separately, incorporated in early work on multimodal generation (André and Rist, 1996) is contradicted by more recent psycholinguistic research, in which gesture and language are increasingly viewed as tightly coupled (Kita and Özyürek, 2003; McNeill, 1985; McNeill and Duncan, 2000), contributing jointly to the composite utterances (Enfield, 2009). This view has also influenced recent work in multimodal NLG. For example, Kopp et al. (2008) use 'multimodal concepts', combining propositional and gestural or perceptual information.

In the case of referring expressions, pointing has been treated as a property, on a par with an object's colour or size. Thus, van der Sluis and Krahmer (2007) propose an algorithm in a graph-based framework (Krahmer et al., 2003) which selects pointing gestures of varying degrees of precision based on their cost when compared to other linguistically realisable features. Similarly, Kranstedt and Wachsmuth (2005) propose an extension of Dale and Reiter's (1995) Incremental Algorithm, which initially considers the possibility of producing an unambiguous pointing gesture. If this fails, a pointing gesture that is less precise may be generated, together with descriptive features of an object.

Both of these approaches assume that the choice of modality in a referring action ultimately hinges on a trade-off between what can be said and what is easiest to produce, a view that has some empirical support (Beun and Cremers, 1998; Bangerter, 2004; Piwek, 2007). On the other hand de Ruiter et al. (2012) found that likelihood of pointing was unaffected by the difficulty of using descriptive features. From a computational perspective, our earlier work (Gatt and Paggio, 2013) also found evidence, based on a machine-learning study on dialogue data, for the co-occurrence of pointing with descriptive (especially spatial) features, suggesting that pointing gestures may be planned in tandem (and not in competition) with these features.

The present paper uses the same corpus data as Gatt and Paggio (2013); however, that paper focused on the relationship between descriptive features (in the spoken part of the utterance) and pointing. In contrast, here we take a much broader view, also addressing the impact of the physical/perceptual features of the objects under discussion, and aspects of the discourse history.

3 Data used in this study



(a) Experiment setup

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

Figure 1: MREDI dialogue setup (reproduced from Gatt and Paggio (2013)).

We use the MREDI (Multimodal REference in DIalogue) corpus (van der Sluis et al., 2008; Gatt and Paggio, 2013), a collection of MapTask-like dialogues (Anderson et al., 1991). Dialogues in MREDI were conducted by dyads consisting of a Director and a Follower. The Director's task was to guide the Follower along a route through a visually shared 'map', located approximately one metre away, directly in front of them, blown up to roughly A0 size. The Director also had a private map on which the route was indicated, while the Follower's private map was used to mark the route as it unfolded in the course of the conversation. Figure 1a displays the basic setup.

There were no restrictions on what interlocutors could say. Participants in the study were told in advance that they could use both speech and gestures, but were not explicitly instructed to point. The maps consisted of collections of shapes of different colours and sizes and were very densely populated (see Figure 1b). Four maps were used in the study: in two of these, landmarks consisted of individual circles or squares, while in the other two they consisted of groups or clusters of five circles or squares (Figure 1b is a group circle map). In the group maps, all elements of a group of five were of the same colour and size.

On each map, there were 18 ‘landmarks’; these were the milestones along the itinerary and were marked on the Director’s private map, but not visible on the large map that constituted the common ground. For example, the landmarks (groups of 5 circles) in Figure 1b are numbered from 1 to 18. Each dyad did all four maps; the order was randomised for each pair of participants. Participants switched roles between one map and another. In addition to the difference between group and individual landmarks, the maps were designed to manipulate a number of independent variables:

1. **Distinguishing Properties (DistProps):** Landmarks 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. This means that different landmarks required different combinations of properties to ensure that they could be unambiguously identified by a description. For example, in Figure 1b, the group marked 17 consists of a landmark where size is the distinguishing feature, since all five circles in the group are small, and the objects in their immediate vicinity are either large or medium-sized. There were equal numbers of landmarks on each map that could be distinguished by colour only, size only, or both.

2. **Prior reference (Discourse):** Some of the landmarks were visited twice in the itinerary; these are indicated using two numbers in Figure 1b. Thus, landmark 8 in this map was also visited later as landmark 13. There were 6 landmarks on each map that were revisited in this way. This is the primary manipulation related to discourse history.

3. **Shift of domain focus (Distance):** Landmarks were located either near to or far away from the previous target. For example, in Figure 1b, landmark 17 and landmark 18 are adjacent (‘near’ condition), but landmark 17 is far from the preceding landmark 16.

In what follows, we use data from 8 dyads. Similar to Gatt and Paggio (2013), we only consider utterances by Directors. 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. Our dataset consists of a total of 2255 such utterances, of which 370 (16.4%) contain a pointing gesture. This is a relatively low proportion of such gestures, compared to some previous studies, such as Beun and Cremers (1998), who found that 48% of referential acts in their task-oriented dialogue corpus included a pointing gesture. However, Beun and Cremers focussed exclusively on first-mention referring expressions. Furthermore, the low proportion of pointing gestures in MREDI may be due to the fact that under our definition, the identification of a landmark may be spread over several turns, with possible interruptions by the Follower. Each such turn constitutes a separate utterance. This raises the likelihood that certain features of the composite utterance, including pointing, will only occur on some of the turns.

3.1 Features

Utterances in MREDI were annotated with the features displayed in Table 1. These codify aspects of the descriptive content of a referential act, as well as the presence or absence of a pointing gesture.

The features originally encoded in the MREDI corpus had frequency values; Gatt and Paggio (2013) used these frequencies in their study. However, for our experiments, we collapsed the features related to descriptive content – hereafter referred to as *descriptive* features – into boolean features. This significantly reduces the feature set and makes the rules acquired in our machine-learning experiments easier

	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	ID	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.

to interpret. Further, it enables us to test our hypothesis that the presence or absence of a *type* of feature (descriptive, physical or discursive) impacts the decision to point. The boolean descriptive features are as follows:

1. **Deixis**: this has the value `true` if the utterance contains a demonstrative pronoun (such as *that*), or a reference to the landmark that identifies it with the previous landmark. Thus, this feature is `true` if $ID > 0$ or $D > 0$ in Table 1;

2. **Locative**: this has the value `true` if the utterance contains any of the spatial properties in Table 1. Thus, the feature is `true` if $AP > 0$ or $RP > 0$ or $FP > 0$ or $DIR > 0$.;

3. **Visual**: this has the value `true` if the utterance contains at least one mention of the landmark's visual properties. This, the feature is `true` if $C > 0$ or $Sh > 0$ or $S > 0$.

In addition to these features, our experiments also made use of the *physical* features (**Distance** and **DistProps**) manipulated as part of the MREDI data collection study (see above), as well as the feature **Discourse**, which encodes prior reference.

Finally, we added a new feature to the dataset, **MapConfl**, which indicates the type of map on which utterances were produced, namely, individual or group circles or squares. This feature was included because the larger size of group landmarks, compared to individuals, may have influenced the decision to point, since groups are more visually salient.

The feature *Gaze* is present whenever a pointing gesture is made; hence, it is not used in the machine learning experiments reported below.

4 Experiments

In our earlier study on the MREDI corpus (Gatt and Paggio, 2013), investigating the relationship between pointing and descriptive features, we found that the latter could indeed be used as predictors of pointing gestures with an accuracy of 0.833 (F-score). The study also concluded that among the descriptive features it was locative properties that were most useful in guiding the decision of whether or not to point, compared to features describing visual characteristics of the objects.

However, in much of the work reviewed in Section 2, especially work arguing in favour of a trade-off in cost between pointing and describing, the occurrence of pointing is made to depend on the physical properties of referents. Therefore, in the present study we want to test whether the occurrence of pointing gestures can be predicted more accurately as a function of (i) the descriptive features that speakers use to refer to landmarks; (ii) the physical/perceptual context in which they are found and (iii) whether or not they have been referred to earlier in the discourse. Furthermore, we want to investigate which combinations of physical and descriptive features provide the best results.

Two sets of experiments were conducted on different versions of the MREDI dataset. The first dataset (referred to as the complete dataset) is the same one used in the Gatt and Paggio (2013) study. It includes all of the 2255 Director's utterances from the eight dyads in the corpus, including those that did not contain any references at all, linguistic or gestural. Such utterances might, for example, be confirmations or feedback produced in the course of the dialogue.

We also report results on a second dataset (referred to as the referential dataset), consisting of all utterances that contain a reference, either using descriptive features, pointing, or both. This dataset consisted of 1542 utterances. Note that the number of utterances with a pointing gesture is still 370 in the pruned dataset.

The task in the experiments was to classify the binary feature *Point*. As mentioned earlier, 370 of these utterances contain a pointing gesture. In other words, there are 370 occurrences of *Point=1*.

All the experiments were run using the Weka tool (Witten and Frank, 2005), which gives access to many different algorithms, using 10-fold cross-validation throughout. In the experiments with the complete dataset, the ZeroR and OneR classifiers were first run on the data to establish a baseline. ZeroR always chooses the most frequent value of the class that is being predicted; in the present case, it consistently classifies all utterances as *Point* = *0*, since the majority of utterances do not containg pointing gestures. OneR identifies a single feature, on the basis of which all classifications are made. On the MREDI data, OneR always assigned *Point* = *0* to all utterances, based on a single rule using the *MapConfl* feature (i.e. the type of map or domain in which the dialogue was being carried out). Note that both baseline classifiers were trained using all features.

Various combinations of descriptive and physical features were then tested using different classifiers in Weka, including NaiveBayes, Support Vector Machines, Maximum Entropy (Logistic in Weka) and the J48 Decision Tree classifier. The present paper will report results for the last two of these, for the following reasons. First, these were the ones which performed best. In addition, the decision trees built by J48 provide an analysis tool to understand how the various features interact, given their transparency; on the other hand, MaxEnt sometimes outperforms J48 and provides a 'ceiling', in addition to the baselines described above.

The strategy used in testing feature combinations was essentially ablative. We tested first using all features, and then compared the performance of the classifiers when they use only descriptive features (Visual, Locative and Deixis), or only Discourse together with the physical features (DistProps and Distance). Omitting descriptive features and using only physical features with Discourse invariably performed near or below baseline (see below). Thus, we experimented with combinations of descriptive features and each physical feature, as well as Discourse, individually.

4.1 Results on the complete dataset

The results for the complete dataset are shown in Table 2 in terms of Precision, Recall and F-measure for each of the classifiers. The top rows display the results using all features, while the baseline results are in the bottom rows. The remaining results for different combinations of features are in descending order of F-score.

Interestingly, using all features – i.e. MapConfl, DistProps, Discourse, Distance, Visual, Locative and Deictic – with or without MapConfl, results in worse overall performance than using a combination of descriptive features (Locative, Deictic and Visual) with Distance. This combination is closely matched for accuracy by the combination involving descriptive features, Distance and DistProps. However, dropping Distance (using only descriptive features and DistProps) results in worse performance.

The addition of Distance and/or DistProps clearly improves the predictive accuracy of a classifier that uses descriptive features. However, the worst combination is found when the descriptive features are excluded. This is in line with the results reported by Gatt and Paggio (2013), who found that features of

Classifier	P	R	F	Features
J48	0.827	0.847	0.832	All
Logistic	0.831	0.854	0.828	All
J48	0.833	0.851	0.838	All - MapConfl
Logistic	0.832	0.851	0.837	All - MapConfl
Logistic	0.839	0.853	0.844	Descriptive + Distance
J48	0.839	0.853	0.844	Descriptive minus Deictic + Distance
Logistic	0.839	0.853	0.844	Descriptive minus Deictic + Distance
J48	0.836	0.851	0.84	Descriptive+DistProps + Distance
J48	0.839	0.853	0.84	Descriptive+Distance
Logistic	0.833	0.851	0.838	Descriptive+DistProps + Distance
J48	0.821	0.847	0.824	Descriptive+DistProps
Logistic	0.809	0.842	0.794	Only Descriptive
Logistic	0.809	0.842	0.794	Descriptive + DistProps
Logistic	0.809	0.842	0.793	Descriptive + Discourse
J48	0.803	0.84	0.787	Only Descriptive
J48	0.795	0.838	0.781	Descriptive + Discourse
J48	0.699	0.836	0.761	Physical + Discourse
Logistic	0.699	0.836	0.761	Physical + and Discourse
ZeroR	0.699	0.836	0.761	All
OneR	0.699	0.836	0.761	All

Table 2: Predicting pointing gestures with different feature combinations in the complete MREDI dataset.

the descriptions produced by speakers were good predictors of pointing.

Adding only DistProps to the descriptive features improved the accuracy of the Logistic classifier somewhat, though it had a greater impact on J48. However, Distance seems to have the greatest impact of the two physical features. Discourse does not appear to play an important role: incorporating this feature does not result in much improvement over using only descriptive features; indeed, in the case of J48, it decreases accuracy.

We also tested one of the best combinations involving descriptive features and Distance but excluding the Deictic feature from the set of descriptive features. This was done because pointing in referential acts is frequently viewed on a par with deictic expressions, insofar as they are both indexical (Bangerter, 2004). This raises the question whether, out of all the descriptive features, Deixis could be considered a somewhat redundant predictor. The results suggest that removing Deixis from the descriptive features does not alter the accuracy of the classifier. We return to the role of Deixis in the discussion in Section 4.3.

4.2 Results on the referential dataset

Exactly the same combinations of features were tested, using 10-fold cross-validation, in separate experiments on the referential dataset. This was done in order to compare the results on a dataset which contains less 'noise', that is, fewer utterances which were non-referential. Such utterances may compromise the predictive validity of certain features, as they inflate the number of utterances in which *Point=0*.

Table 3 contains the results obtained on the reduced dataset. The accuracy is in general lower due to the fact that predicting the absence of pointing is easier in the complete dataset, where many utterances contain no reference at all, descriptive or gestural.

Contrary to the findings on the complete dataset, using the complete set of features as predictors of pointing gives slightly better results than using either descriptive or physical features alone, at least in

Classifier	P	R	F	Features
J48	0.783	0.799	0.785	All - MapConfl
Logistic	0.726	0.764	0.679	All - MapConfl
J48	0.774	0.793	0.776	All
Logistic	0.704	0.760	0.681	All
J48	0.781	0.797	0.784	Descriptive + DistProps + distance
J48	0.766	0.785	0.770	Descriptive + DistProps
J48	0.748	0.777	0.745	Descriptive + Distance
J48	0.758	0.783	0.744	Only descriptive
J48	0.774	0.788	0.740	Descriptive + Discourse
Logistic	0.720	0.762	0.675	Descriptive + DistProps + distance
Logistic	0.688	0.759	0.662	Descriptive + Discourse
Logistic	0.699	0.760	0.661	Descriptive + DistProps
Logistic	0.759	0.761	0.660	Descriptive + Distance
J48	0.577	0.760	0.656	Only physical + Discourse
Logistic	0.577	0.760	0.656	Only descriptive
Logistic	0.577	0.760	0.656	Only physical + Discourse
ZeroR	0.577	0.76	0.656	All
ONeR	0.577	0.76	0.656	All

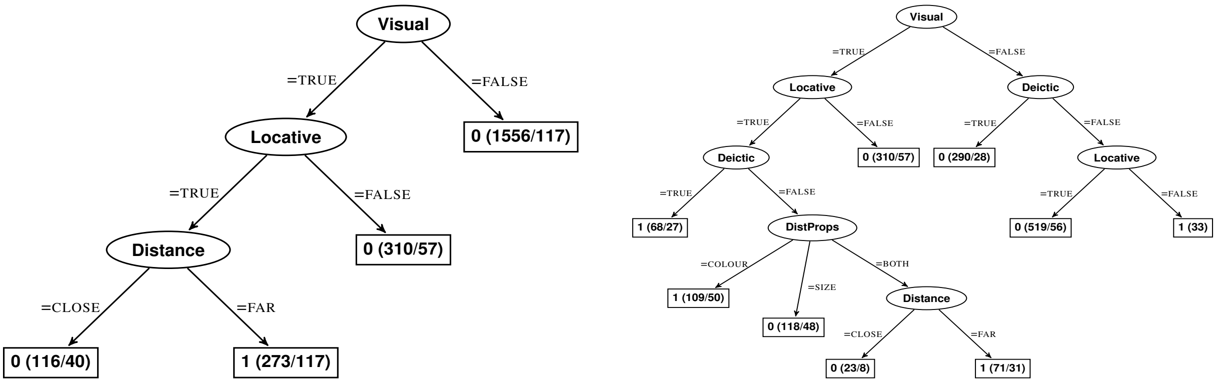
Table 3: Predicting pointing gestures with different feature combinations in the referential MREDI dataset.

the case of the decision tree classifier. This combination also exceeds the combination of descriptives, DistProps and Distance, though only marginally. However, this does remain the next best combination for J48, consistent with the results on the complete dataset. However, this combination performs quite badly in the case of the Logistic classifier.

The fact that using all features performs better this time is probably due to the fact that there are fewer non-referential utterances in this dataset. Once again, the role of Discourse seems marginal.

4.3 Analysis and discussion

Figure 2 shows the decision trees built by J48 for the two datasets when descriptive features are used together with *DistProps* and *Distance*.



(a) Complete MREDI dataset

(b) Referential MREDI dataset

Figure 2: J48 Decision trees from the complete and referential datasets

Our main findings can be summarised as follows. First, descriptive features play an important role in the prediction of pointing; this replicates previous observations (Gatt and Paggio, 2013). Second, and more importantly, the prediction accuracy improves when physical features, representing aspects of the visual/perceptual context, are taken into account. This is especially true of Distance, suggesting that a sizeable shift of perceptual focus, from one landmark to another further away, motivates a pointing gesture, as shown in both trees in Figure 2. Once again, it is worth comparing this to the results of Beun and Cremers (1998), who find that shifts of perceptual focus play a role in increasing the amount of (descrpitive) information speakers include in a referring expression. However, they find no impact of focus shifts on pointing gestures; our results, by contrast, suggest that such shifts do play a role.

There are a number of striking features in the trees in the figure. First, the descriptive feature Locative plays a crucial role. All cases of pointing involve the presence of a Locative, with one exception: on the referential dataset (Figure 2b), in case no Visual, Deictic or Locative features are used, the tree predicts a pointing gesture. However, this case covers a very small number of instances (33), with 0% error rate. All of these turn out to be utterances where there is no descriptive reference at all and speakers rely exclusively on pointing. Example (3) below is typical of these.

(3) D: And a slightly bigger green to the right of that
M: M-hm
D: In the center of those like pack
M: Yeah
D: is number 9. [+pointing]

Clearly, these are cases in which the pointing gesture occurs as part of an extended sequence of utterances which jointly identify a landmark. Descriptive features have already been uttered; the pointing comes at the very end. In summary, the one case where Locatives don't feature in predicting a pointing gesture turns out to be a rather special case.

A second striking aspect of the trees is that while Deixis plays a predictive role in the tree based on the referential dataset, it doesn't in the case of the complete dataset. This is interesting in view of the relationship that has often been noted between referential pointing gestures and deictic expressions (Bangerter, 2004). Note, however, that there is no inconsistency between the two trees: the single path through the tree in Figure 2a that results in pointing is subsumed by the path in Figure 2b which specifies in addition that Deixis should be `false`, and DistProps should have the value `colour`. This still leaves open the question why Deixis plays no role in the full dataset, despite being included as part of the descriptive features that resulted in this tree. Indeed, we have already shown that, among the descriptive features, Deixis doesn't contribute much predictive power on the full dataset (see Section 4.1).

One possibility is that Deixis is generally under-represented in the corpus. However, there are proportionately fewer utterances in the full dataset containing Deixis (20%), compared to the referential dataset (30%). Furthermore, it may be partially dependent on the Locative features. There may be a priori reasons to assume this as a working hypothesis: Deixis anchors parts of the speech signal to physical properties of the common ground, potentially making it redundant with respect to location (which has already specified the relevant physical/spatial features of the common ground).

	Complete Dataset		**Referential Dataset**	
Locative	*Deictic*		*Deictic*	
	`false`	`true`	`false`	`true`
`false`	74	26	42	58
`true`	88	12	88	12
overall	80	20	70	30

Table 4: Deictic features (D and ID) relative to Locatives. All figures are percentages.

Table 4 displays the distribution of Deictic expressions with respect to Locatives, that is, the proportion of utterances containing a Deictic expression as a function of whether the utterances also contain a Locative expression. The tables shows proportions both for the full and the referential dataset.

Note that when Deictics are used, it is mostly in the absence of a Locative expression. A chi-square test of independence suggests that the frequency of use of Locative and Deictic expressions are not independent (complete dataset: $\chi^2_1 = 63.044, p < .001$; referential: $\chi^2 = 358.21, p < .001$). However, there is a higher proportion of Deictic expressions in the referential dataset (30% overall); this may account for the use of this feature in the decision tree for this dataset (it is more informative). Crucially, the trend in the use of deictic expressions is reversed in the two datasets: when Locative is false on the referential dataset, most utterances involve a deictic expression; the reverse is true on the complete dataset.

There is one path through the tree in Figure 2b which seems to contradict the hypothesis that deictic expressions are used in the absence of locatives. There are 68 cases where pointing is used when both Deictic and Locative are true. One possibility is that this is caused by our having defined the Deictic feature as true whenever there is an actual deictic expression (variable D in Table 1; e.g. *those squares*) or an identity expression (variable ID; e.g. *the same one we saw*). To investigate this further, Table 5 shows a breakdown of the frequencies of the presence or absence of a locative expression, as a function of whether a true deictic (D) or an identity expression (ID) is used in an utterance. Once again, proportions are displayed for both datasets.

	Complete dataset		**Referential dataset**	
True Deictic	*Locative*		*Locative*	
	false	true	false	true
false	54	46	26	74
true	78	22	76	24

(a) True deictic expressions (D)

	Complete dataset		**Referential dataset**	
Identity	*Locative*		*Locative*	
	false	true	false	true
false	56	44	31	69
true	68	32	67	33

(b) Identity expressions (ID)

Table 5: Identity (ID) and actual Deictic (D) expressions relative to Locatives. All figures are percentages.

There are two observations that stem from these proportions: First, in line with our earlier observations, there is a greater proportion of true deictic (D) expressions in utterances that contain no locative expression. For example, 78% of utterances in the complete dataset that have no locatives contain a deictic; the corresponding figure in the referential dataset is 76%. The same pattern holds for identity (ID) expressions. Second, out of the utterances that do not contain a locative, the proportion containing a true deictic (D) is greater than the proportion containing an identity expression (ID). This may explain the apparent exception – represented by the path in Figure 2b – to our generalisation that locatives and deictics are redundant with respect to each other, and locatives tend to be avoided if deictics are used. The explanation may lie in the conflation, in the boolean Deictic feature used in our experiments, of true deictics and identity expressions. The path in the decision tree where both Locative and Deictic are true may be accounting for utterances in which an identity expression is used, rather than a true deictic.

5 Conclusions and future work

This paper addressed the question of when pointing gestures should be generated, as a function of the features a speaker uses to identify a referent, as well as the features of the context in which an utterance is being produced. The best predictors of pointing are descriptive features, especially locatives, and features of the perceptual context, especially distance from the last referent. The latter is a marker of a shift of perceptual focus, akin to the focus shifts identified by Beun and Cremers (1998). Our study also sheds light on the relationship between pointing and the use of deictic expressions, suggesting that, while the two are often used together, deictics tend to be used more in the absence of locative features.

We also note some limitations of our methodology. Inspection of the results in Tables 2 and 3 shows that the best performing classifiers, though they exceed baselines, do not do so by a wide margin. We believe that one of the main reasons for this is the relative scarcity of pointing gestures in our dataset (as discussed in Section 3), which may have resulted in a sizeable subset of utterances where pointing was relatively straightforward to predict (e.g. based on one feature, as in the OneR baseline classifier). This is a limitation we intend to investigate in future work, through a more diverse dataset where pointing features more strongly. In addition, it is worth noting that the ablative testing reported here does suggest that certain features play a greater role in determining when speakers choose to point.

Our work addresses an important question in Natural Language Generation systems that seek to generate multimodal referring acts, namely, how pointing and describing should be combined and when. In future work, we intend to extend this research in two ways: first, by extending our focus to incorporate the interactive features of a dialogue and their impact on referential success; and second, by focusing on other domains with a view to testing the generalisability of the results.

Acknowledgements

Thanks to Ielka van der Sluis, Adrian Bangerter and Paul Piwek, who collaborated on the development of the MREDI corpus. Thanks to the anonymous reviewers of COLING 2014 for helpful comments.

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.

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

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. 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.

N.J. Enfield. 2009. *The Anatomy of Meaning: Speech, Gesture and Composite Utterances*. Cambridge University Press, Cambridge.

A. Gatt and P. Paggio. 2013. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *Proceedings of the 14th European Workshop on Natural Language Generation (*ENLG'13*).*

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 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).*

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.

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.