

The Impact of Visual Context on the Content of Referring Expressions

Jette Viethen^{1,2}

h.a.e.viethen@uvt.nl

¹TiCC

University of Tilburg
Tilburg, The Netherlands

Robert Dale²

robert.dale@mq.edu.au

²Centre for Language Technology

Macquarie University
Sydney, Australia

Markus Guhe³

m.guhe@ed.ac.uk

³School of Informatics

University of Edinburgh
Edinburgh, UK

Abstract

Traditional approaches to referring expression generation (REG) have taken as a fundamental requirement the need to distinguish the intended referent from other entities in the context. It seems obvious that this should be a necessary condition for successful reference; but we suggest that a number of recent investigations cast doubt on the significance of this aspect of reference. In the present paper, we look at the role of visual context in determining the content of a referring expression, and come to the conclusion that, at least in the referential scenarios underlying our data, visual context appears *not* to be a major factor in content determination for reference. We discuss the implications of this surprising finding.

1 Introduction

Traditional approaches to referring expression generation are based on the idea of distinguishing the intended referent from the other entities in the context (Dale and Reiter, 1995; Gardent, 2002; Krahmer and Theune, 2002; Krahmer et al., 2003; Gatt and van Deemter, 2006). The task is generally characterised as involving the construction of a *distinguishing description* consisting of those attributes of the intended referent that distinguish it from the other entities with which it might be confused; building a referring expression thus requires us to have an appropriate formalisation of the notion of *context*. Earlier work (for example, (Dale, 1989)) took its cue from work on discourse structure (in particular, (Grosz and Sidner, 1986)), and defined the context in terms of the set of discourse-accessible referents; more recent work has tended to focus on visual scenes (for example, (Viethen and Dale, 2006; Gatt et al., 2008; Gatt et al., 2009)), with the context being defined as the set of all the objects in the scene.

Most of the early approaches to REG (Dale, 1989; Dale and Haddock, 1991; Dale and Reiter, 1995; Krahmer et al., 2003) were proposed without the support of rigorous empirical testing. Probably the most fundamental shift in the field in the last five years has been the move towards

the development of algorithms that attempt to replicate corpora of human-produced referring expressions. This work has only really become possible with the advent of a number of publicly-available corpora of human-produced referring expressions collected under controlled circumstances: these include the TUNA Corpus (van der Sluis et al., 2006), the Drawer Corpus (Viethen and Dale, 2006), and the GRE3D3 and GRE3D7 Corpora (Viethen and Dale, 2008; Viethen and Dale, 2011). All of these corpora contain descriptions of target referents using a small number of attributes in simple visual scenes containing only a very small number of distractor objects. The descriptions in all these cases were elicited in isolation, with no preceding discourse: the reference task they represent has sometimes been called ‘one-shot reference’. So there is no *discourse* context that provides a set of potential distractors, but there is a *visual* context of potential distractors.

The idea that the process of constructing a reference to an object in a visual scene needs to take account of the other entities in that scene in order to ensure that the reference is successful seems so obvious that it might be thought ridiculous to doubt it. However, our exploration of a dataset that contains referring expressions for objects in visual scenes of somewhat greater complexity and involving dialogic discourse calls this fundamental assumption into question.

In (Viethen et al., 2011), we presented a machine-learning approach to REG, and distinguished two main kinds of features that might play a role in subsequent reference: ‘traditional’ REG features, which are concerned with distinguishing the intended referent from visual and discourse distractors; and ‘alignment’ features, representing aspects of the discourse history (Clark and Wilkes-Gibbs, 1986; Pickering and Garrod, 2004). We used feature ablation in a decision tree approach to investigate the role of the traditional features, and found that the impact of these features was negligible compared to that of the alignment features. The bad performance of these features caused us to ask whether the method of determining

the visual distractors that were taken into account was to be blamed. In the present paper, we explore this question by trying out two different ways of determining the set of visual distractors and by varying the size of this set.

In Section 2 we provide some background by situating the investigation presented here with respect to the literature. In Section 3, we describe the corpus we work with, and in Section 4, we describe our machine-learning framework for exploring the data this corpus provides. In Section 5, we present the results of some experiments that attempt to determine the role of visual context in REG, and in Section 6 we draw some conclusions.

2 Background

Some of the earliest work in REG (for example, (Dale, 1989)) adopted what we might think of as an ‘extreme rationalist’ characterisation of the task: build a description that has no more and no less information than is required to distinguish the intended referent (a *minimal distinguishing description*).

It was soon recognised that this was not a good characterisation of what people did, in particular because human-produced descriptions are often over-specified, rather than being minimal in the sense just described. The incremental algorithm (IA; (Dale and Reiter, 1995)) diluted the extreme position with the acknowledgement that something akin to habit also played a role in REG: the basic idea here was that, on the basis of experience, people learn ‘preference orders’ for properties that tend to work well, and when faced with the need to create a new description, they use these preference orders to guide the search for an appropriate description. The IA still hung on to the need to build a distinguishing description, but the preference order mechanism meant that some descriptions might be longer than necessary, containing redundant information.

In (Dale and Viethen, 2010), we proposed a further weakening of the traditional model, suggesting that attributes in a referring expression might be chosen independently, rather in a fashion whereby each depends on the attributes previously chosen (a characteristic of earlier algorithms that we refer to as *serial dependency*). But even this attribute-centric model takes the view that the discriminatory power of the individual attributes plays a role in decision-making. The requirement that we should take account of the context in determining how to refer to something has thus been kept more or less centre-stage in computational work through the last 20 years or so.

Meanwhile, work in psycholinguistics has explored the idea that quite orthogonal factors are at play in choosing the content of descriptions. Starting with the early work of Carroll (1980), a distinct strand of research has explored how a speaker’s form of reference to an entity is

impacted by the way that entity has been previously referred to in the discourse or dialogue. The general idea behind what we will call the *alignment approach* is that a conversational participant will often adopt the same semantic, syntactic and lexical alternatives as the other party in a dialogue. This perspective is most strongly associated with the work of Pickering and Garrod (2004). With respect to reference in particular, speakers are said to form *conceptual pacts* in their use of language (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). The implication of much of this work is that one speaker introduces an entity by means of some description, and then (perhaps after some negotiation) both conversational participants share this form of reference, or a form of reference derived from it, when they subsequently refer to that entity. Recent work by Goudbeek and Krahmer (2010) supports the view that subconscious alignment does indeed take place at the level of content selection for referring expressions: the participants in their study were more likely to use a dispreferred attribute to describe a target referent if this attribute had recently been used in a description by a confederate.

One way of characterising these developments is that, on the one hand, the original very precise and somewhat rigid computational approaches to REG have been progressively weakened in the face of real human data; and on the other hand, work in a distinct discipline has offered a quite separate view of how reference works. Of course, these two broad approaches may not be incompatible. The truth may lie ‘in-between’, involving insights and ideas from both ways of thinking about the problem. In the present paper we aim to put one of the remaining fundamental tenets of the computational approaches to the test: does visual context really matter when we construct a referring expression?

3 Referring Expressions in the iMAP Corpus

The iMAP Corpus (Louwerse et al., 2007) is a collection of 256 dialogues between 32 participant-pairs who contributed 8 dialogues each. Both participants had a map of the same environment, but one participant’s map showed a route winding its way between the landmarks on the map (see Figure 1 for examples). The task was for this participant, the instruction giver (IG), to describe the route in such a way that their partner, the instruction follower (IF), could draw it onto their map; this was complicated by some discrepancies between the two maps, such as missing landmarks, the unavailability of colour in some regions due to ink stains, and small differences between some landmarks. Note that the maps contain a relatively large number of objects compared to the visual stimuli used in other REG corpora.

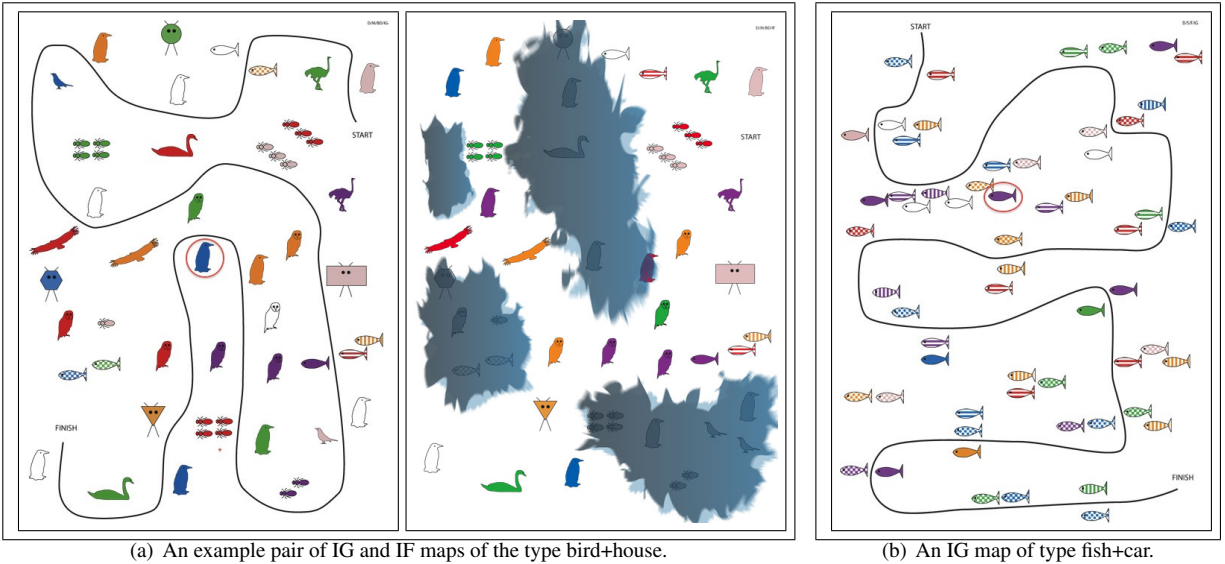


Figure 1: Three example maps.

There are eight types of landmarks, grouped into pairs of one animate and one inanimate type each: alien+traffic sign, bird+house, fish+car, and bugs+trees. Each of these pairs defines a map type, which contains landmarks which are mostly of one of the two types of the pair. Half of the maps contain a few landmarks of types other than the main type; for example, a bird+house map contains mostly birds or houses, but might also contain a small number of other landmarks. The maps in Figure 1(a) are bird+house maps containing mainly birds with a few landmarks of other types mixed in, and the map in Figure 1(b) is an unmixed fish-car map for the IG, containing only fish landmarks. Note the high density of landmarks on the map in Figure 1(b) compared to those in Figure 1(a) (each cluster of same-coloured bugs on the bird maps counts as a single landmark). Overall there are 32 maps, which differ by the map type (four levels), the animatedness of the landmark types (two levels, e.g. fish vs. cars), the mixedness of the landmark types (two levels: only the main landmark type or also a few landmarks of different types), and the shape of the ink blots on the IF's map (two levels: one large blot or several smaller ones).

Apart from their type, the landmarks differ in colour, and one other attribute, which is different for each type of landmark. For example, there are different *kinds* of birds and houses (eagle, ostrich, penguin, ...; church, castle, ...); fish and cars differ by their *patterns* (dotted, checkered, plain, ...), aliens and traffic signs have different *shapes* (circular, hexagonal, ...), and bugs and trees appear in small clusters of differing *numbers*. In addition to these three inherent attributes of the landmarks, par-

ticipants used spatial relations to other items on the map. Each of the 34,403 referring expressions in the corpus is annotated with the semantic values of the attributes that it contains. This collection of annotations forms the basic data we use in our experiments.

We removed from the data all referring expressions that made reference to more than one landmark and those—in particular, pronouns—that did not contain any of the four main landmark attributes, type, colour, relation, or the landmark's other distinguishing attribute. However, all filtered expressions are taken into account in the computation of the features for the machine learner. The final data set contains 22,727 referring expressions, of which 6,369 are initial references and 16,358 are subsequent references.

We can think of each referring expression as being realised from a *content pattern*: this is the collection of attributes that are used in that description. The attributes can be derived from the property-level annotation given in the corpus. So, for example, if a particular reference appears as the noun phrase *the blue penguin*, annotated semantically as $\langle \text{blue, penguin} \rangle$, then the corresponding content pattern is $\langle \text{colour, kind} \rangle$. Our aim is to replicate the content pattern of each referring expression in the corpus. Table 1 lists the 15 content patterns that occur in our data set in order of frequency.

The high frequency of the $\langle \text{other} \rangle$ pattern is in part due to the annotation of the kind of birds and houses as other, which could also be argued to be a more fine-grained type attribute. We accepted this annotation as it was provided in the corpus, but we may alter it in future studies.

Content Pattern	Count	Proportion
<other>	7561	33.27%
<other, type>	5975	26.29%
<other, colour>	2364	10.40%
<other, colour, type>	1954	8.60%
<colour>	1029	4.53%
<relation>	796	3.50%
<other, relation>	738	3.25%
<type>	662	2.91%
<colour, type>	596	2.62%
<other, relation, type>	463	2.04%
<relation, type>	262	1.15%
<other, colour, relation>	124	0.55%
<colour, relation>	101	0.44%
<other, colour, relation, type>	82	0.36%
<colour, relation, type>	20	0.09%
total	22,727	

Table 1: The 15 different content patterns that occur in our data and their frequencies.

4 A Machine Learning Approach to Content Determination

The number of factors that can be hypothesised as having an impact on the form of a referring expression in a dialogic setting associated with a visual domain is very large. Attempting to incorporate all of these factors into parameters for a rule-based system, and then experimenting with different settings for these parameters, is prohibitively complex. Instead, we here capture a wide range of factors as features that can be used by a machine learning algorithm to automatically induce from the data a classifier that predicts for a given set of feature values the attributes that should be used in a referring expression.

The features we extracted from the data set are outlined in Tables 2–4.¹ They fall into a number of subsets. **Map** features capture design characteristics of the map-pair the current dialogue is about; **Speaker** features capture the identity and role of the participants; and **LMprop** features capture the inherent visual properties of the target referent. The **TradREG** features allow the machine learner to capture factors that the traditional computational approaches to referring expression generation take account of. Of particular interest for our present considerations are the **Visual TradREG** features, which represent knowledge about the visual context. **Alignment** features capture factors that we would expect to play a role in the psycholinguistic models of alignment and conceptual pacts. When we refer to the complete feature set, we use the abbreviation **allF**.

¹In these tables, *att* is an abbreviatory variable that is instantiated once for each of the four attributes type, colour, relation, and the other distinguishing attribute of the landmark. The abbreviation LM stands for landmark.

Map Features	
Main_Map_type	most frequent type of LM on this map
Main_Map_other	other attribute if the most frequent type of LM
Mixedness	are other LM types present on this map?
Ink_Orderliness	shape of the ink blot(s) on the IF’s map
LMprop Features	
other_Att	type of the other attribute of the target
[att]_Value	value for each <i>att</i> of target
[att]_Difference	was <i>att</i> of target different between the two maps?
Missing	was target missing one of the maps?
Inked_Out	was target inked]_out on the IG’s map?
Speaker Features	
Dyad_ID	ID of the pair of participant-pair
Speaker_ID	ID of the person who uttered this RE
Speaker_Role	was the speaker the IG or the IF?

Table 2: The Map, LMprop and Speaker feature sets.

Visual TradREG Features	
Count_Vis_Distractors	number of visual distractors
Prop_Vis_Same_[att]	proportion of visual distractors with same <i>att</i>
Dist_Closest	distance to the closest visual distractor
Closest_Same_[att]	has the closest distractor the same <i>att</i> ?
Dist_Closest_Same_[att]	distance to the closest distractor of same <i>att</i> as target
Cl_Same_type_Same_[att]	has the closest distractor of the same type also the same <i>att</i> ?
Discourse TradREG Features	
Count_Intervening_LMs	number of other LMs mentioned since the last mention of the target
Prop_Intervening_[att]	proportion of intervening LMs for which <i>att</i> was used AND which have the same <i>att</i> as target

Table 3: The TradREG feature set.

For our experiments, we use the Weka Toolkit (Witten and Frank, 2005) to learn one decision tree for each of the four attributes which decides whether or not to include that attribute. We then combine the attributes for which a positive decision was made into a content pattern that can be compared to the content pattern found in the corpus for the same instance.²

In (Viethen et al., 2011) we showed that dropping the complete TradREG feature set from allF does not decrease the performance of this model on subsequent reference. The relevant numbers from that experiment are shown in italics in the first two lines of Table 5.

One question this kind of work raises is: just what gets included in the visual context? Considering that most of the TradREG features depend on the visual context, it might be possible that the lack of impact of this feature set was due to the size of the visual context having been chosen incorrectly. A second consideration is that the TradREG features might have more of an impact on

²We also tried an alternative approach of learning the whole content pattern at once with very similar results, which we do not report here due to space limitations.

Alignment Features – Recency	
Last_Men.Speaker.Same	who made the last mention of target?
Last_Mention_[att]	was <i>att</i> used in the last mention of target?
Dist_Last_Mention_Utts	distance to the last mention of target in utterances
Dist_Last_Mention_REs	distance to the last mention of target in REs
Dist_Last_[att].LM_Utts	distance in utterances to last use of <i>att</i> for target
Dist_Last_[att].LM_REs	distance in REs to last use of <i>att</i> for target
Dist_Last_[att].Dial_Utts	distance in utterances to last use of <i>att</i>
Dist_Last_[att].Dial_REs	distance in REs to last use of <i>att</i>
Dist_Last_RE_Utts	distance to last RE in utterances
Last_RE_[att]	was <i>att</i> mentioned in the last RE?
Alignment Features – Frequency	
Count_[att].Dial	how often has <i>att</i> been used in the dialogue?
Count_[att].LM	how often has <i>att</i> been used for target?
Quartile	quartile of the dialogue the RE was uttered in
Dial_No	number of dialogues already completed +1
Mention_No	number of previous mentions of target +1

Table 4: The Alignment feature set.

initial reference than on the subsequent referring expressions that were at focus in our previous work. We explore these possibilities next.

5 The Effects of Variation in Visual Context

In (Viethen et al., 2011), the size of the visual context was set for each map type in such a way that each landmark on any map of that type would have six distractors on average. We will refer to this way of setting the visual context size as *average-6*.

Because we are here particularly interested in the performance of the features that depend on the visual context (i.e., the Visual TradREG features), we performed two more ablation steps, in which we separately excluded only the Visual TradREG features and the Discourse TradREG features for both subsequent and initial references. Table 5 confirms that, using the *average-6* method to determine the visual context, the Visual TradREG features have no significant effect for either subsequent or initial referring expressions on the Accuracy with which the model replicates the referring expressions in our corpus. Perhaps surprisingly, this is true not only for subsequent reference, but also for initial reference, where one might expect that distinguishing from the visual context would be of more importance.

Considering the difference in density and uniformity of landmarks on the different types of maps (compare Figure 1(a) with 42 diversely shaped landmarks in the IG map to Figure 1(b) with 59 uniformly shaped landmarks), we wondered whether the *average-6* method of setting the visual context might be too inflexible. For ex-

	all	initial	subseq.
allF	61.5%	68.6%	58.8%
allF – TradREG	61.3%	69.4%	58.2%
allF – Discourse TradREG	61.3%	68.6%	58.4%
allF – Visual TradREG	61.6%	69.4%	58.5%
no of REs	22727	6369	16358

Table 5: Ablation of Discourse and Visual TradREG features using *average-6* to determine the visual context. Performance is measured in percentage of perfect matches. Numbers in italics were previously reported in (Viethen et al., 2011).

ample, one might hypothesise that fewer surrounding objects might get taken into account in describing the blue penguin marked by a circle in the left map in Figure 1(a) than in describing the purple fish marked by a circle in Figure 1(b).

We therefore split our data into four sets according to the four different map types and tried out a range of different visual context sizes for each type separately. Two different ways of determining the visual context might be at play. One possibility is that people might indeed be taking into account (roughly) the same number of surrounding objects for each landmark, while this number might be different for different map types due to their different landmark densities. We call this the *count* method of determining the visual context. Alternatively, one might draw an imaginary circle around each landmark, and consider all objects whose centres fall within the radius of this circle to be distractors. We call this the *distance* method of determining the visual context.

In order to explore whether there is one ‘correct’ size of visual context for each map type, we tried all distances from 0 to 675 pixels in 15 pixel steps (each map is 488 × 675 pixels) and all possible distractor counts from 0 to 61 (the maximum number of landmarks on the most dense map pair is 61). If the bad performance of the Visual TradREG features so far was indeed due to the visual context being too inflexible or set incorrectly, we would expect to find at least one visual context size for each map type that outperforms all others. There should also be a peak of performance around that size, with the performance falling if the size grows or shrinks from the ideal size (if the visual context is set too small, we might expect to see references containing too many attributes; if the visual context is set too large, we might expect to see references with too few attributes).

We trained the decision trees on 80% of the data for each map type and tested on the remaining 20%. The training–test splits were stratified for the content patterns of the referring expressions, the Speaker_IDs of the participants who produced the expressions, and the Quartiles of the dialogue in which the references occurred. Table 6

map type	train	test	total
alien+sign	4,425	967	5,392
fish+car	4,021	813	4,834
bird+house	5,492	1,264	6,756
tree+bug	4,703	1,042	5,745
total	18,641	4,086	22,727

Table 6: Sizes of the training and test sets for the different map types.

maptype	best sizes	all REs	best sizes	initial REs	best sizes	subseq. REs
alien+sign	43	63.5%	5	68.3%	43	62.5%
fish+car	44, 46	59.2%	43	60.6%	13	59.0%
house+bird	3, 22	72.6%	22	75.6%	13, 19, 28	71.8%
trees+bugs	3	70.5%	0, 1, 3, 11, 12	74.8%	33	68.4%
weighted average		67.1%		71.1%		65.9%
all maps average-6		61.5%		68.6%		58.8%

Table 7: Maximum possible Accuracy using all features achieved by choosing the best performing visual context by the *count* method for each map type, compared to the performance of the *average-6* visual contexts.

shows the sizes of the four different training–test splits.

Table 7 shows that if we choose the best performing count of distractors for each map type, the overall performance (weighted average over all map types) does indeed improve over the old *average-6* method of choosing the visual context. Table 8 shows the same results for the *distance* method of determining the visual context. (For both methods $p \ll 0.01$, using the χ^2 statistic with $df = 1$ for all, initial, and subsequent references.)

However, Figures 2 to 5 demonstrate that there is no consistent effect of the size of the visual context on the performance of our model using the *number* method of setting visual context sizes. None of the graphs show a clear performance peak around one particular visual context; instead, performance oscillates in a fairly narrow percentage band both when using all features and when using only the Visual TradREG features that are directly impacted by the visual context. For most map types it becomes clear that even a model using only the features that are not affected by the visual context (the flat lines labelled *noVisualTrad*) outperforms allF with many of the settings for visual context size. This means that, unless we are certain that we are using the best performing setting for visual context, using the Visual TradREG features is risky, as choosing the wrong visual context can easily lead to a worse match with human behaviour.

maptype	best sizes	all REs	best sizes	initial REs	best sizes	subsequ. REs
alien+sign	90, 105	59.5%	90	65.1%	240, 285	57.9%
fish+car	75	57.3%	75, 180	62.4%	75	55.9%
house+bird	150	73.3%	300, 540-675	74.8%	480	73.4%
trees+bugs	210	70.4%	585, 660, 675	76.6%	210, 420, 525	67.2%
weighted average		65.9%		70.9%		64.3%
all maps average-6		61.5%		68.6%		58.8%

Table 8: Maximum possible Accuracy using all features achieved by choosing the best performing visual context by the *distance* method for each map type, compared to the performance of the *average-6* visual contexts.

For space reasons we do not show all four graphs for the *distance* method. However, Figure 6 shows the performance for all map types when using all feature sets. Again, the performance oscillates as the size of the visual context varies, rather than showing a real peak around an ideal context size.

Although the performance of the overall system can be increased over the old *average-6* method by setting the visual context to a map type-specific optimum, these results show that this increase is somewhat a matter of luck. Short of trying out (almost) all possible sizes of the visual context, as we did here, there is no systematic way in which to determine the size of the visual context that gives the best performance; and by using features dependent on the visual context one might just as likely hit on a visual context that decreases performance. The oscillations in the graphs in Figures 2 to 6 indicate that it is unlikely that people are taking the visual content into account in the way that our model suggests.

6 Discussion

In this paper we have put forward what might be considered a rather heretical position: that during the construction of a referring expression, contrary to what is assumed by much work in the field, a speaker does not seem to take account of the visual context of reference. Using a collection of human-produced referring expressions of landmarks on moderately complex maps, we have shown that there is no principled way in which to determine a visual context that might make a significant difference to the ability of a machine-learned algorithm to replicate the human data. The implication of this would seem to be that humans generate referring expressions with little regard for the visual context, or at least that the role of visual context is masked by other factors (such as alignment) that play a bigger role. So, we might conclude that

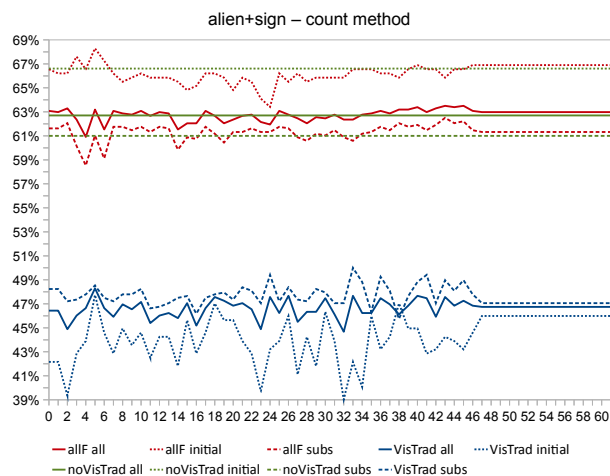


Figure 2: Accuracy for different visual contexts (determined by the *count* method) for the alien+sign maps.

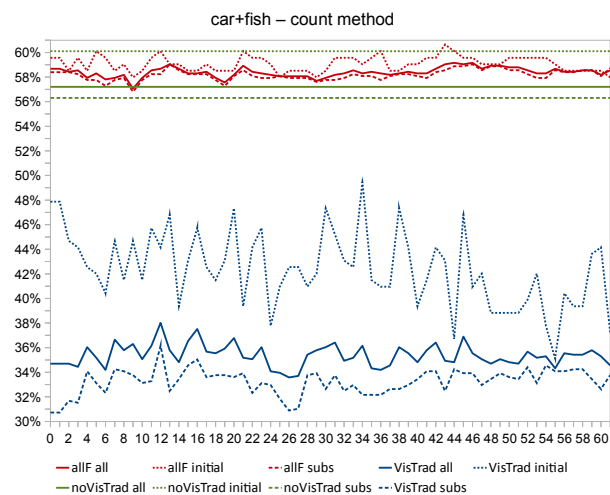


Figure 3: Accuracy for different visual contexts (determined by the *count* method) for the fish+car maps .

the view that reference is about deliberately constructing distinguishing descriptions should be considered suspect.

It could be argued that this is a somewhat plausible position if we look only at *subsequent* reference as we did in (Viethen et al., 2011): once an entity has been introduced into the discourse, perhaps how it is referred to subsequently depends more on the preceding discourse than it does on the visual context at the time of reference. Indeed, once an entity has been referred to, the description that has been constructed ‘factors in’ the visual context, and so any subsequent reference to that entity does not require re-computation of the description; referring to the entity in the way that it was referred to before should still do the job (unless, of course, the context has changed in some relevant way). Such a model has the twin appeals of being both more computationally efficient, and consistent

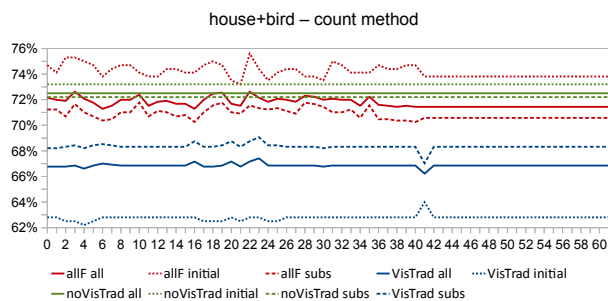


Figure 4: Accuracy for different visual contexts (determined by the *count* method) for the bird+house maps.

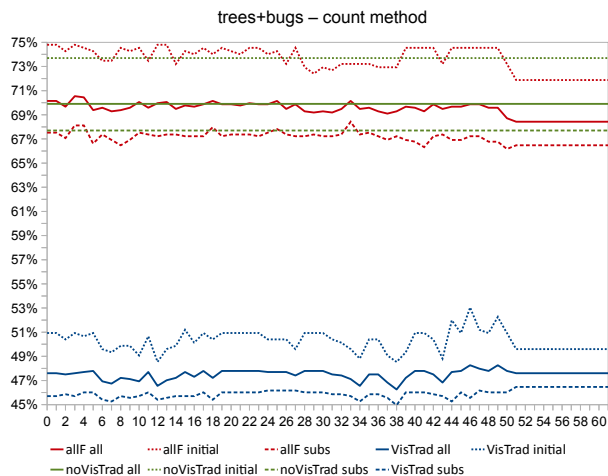


Figure 5: Accuracy for different visual contexts (determined by the *count* method) for the bugs+trees maps.

with explanations based on the alignment approach.

But surely, we would want to say, context must still be taken account of when constructing an initial reference; and if the context is a visual one, then that first reference constructed needs to distinguish the intended referent from the other entities in the scene. Surprisingly, even here, our experimental results support the view that visual context doesn’t matter.

So what’s going on? Intuition suggests that, in real world scenes, we *do* take account of the distinguishing ability of our referring expressions; when we describe an intended referent, we do not do so blindly without considering whether the referring expression might be confusing or ambiguous. But our data suggests, at least in the scenarios we have looked at, that this is not the case.

One possible explanation is that neither of the two ways of determining the visual context that we tried out in our experiments accurately models the visual context that the speakers in our corpus take into account. Firstly, while acknowledging that there are differences between the different types of maps that might influence the number of distractors to be taken into account, we still kept

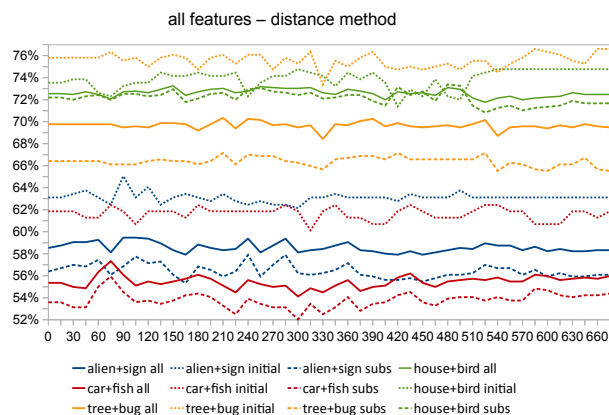


Figure 6: Accuracy for different visual contexts determined by the *distance* method for all map types.

the size of the visual context constant for all landmarks on a given map. It is conceivable that this is still too simplistic an assumption and that distractor numbers have to be determined on a landmark-by-landmark basis instead. For instance, it is likely that, at least for the IG, the course of the path influences the shape of the visual context, with objects along the path being more likely to be taken into account than those further away. This is a consideration that was taken into account to some extent by Guhe (2007; 2009). Similarly, what counts as the visual context is probably influenced by the linguistic context as well. For example, in uttering as well as resolving an instruction such as *go left until you get to the red alien*, the red alien has to be distinguished mostly from objects to its right and not so much from anything that lies beyond it to its left.

To explore these kinds of hypotheses, a lot more preparatory work would be necessary. The dialogues would need to be annotated with information about the point on the path that the IG and IF have reached, and with possibly relevant information in the dialogue context. However, to obtain a more definite answer to the question of which landmarks are taken into account when people refer in dialogue, we will ultimately have to look beyond the text of the dialogue transcriptions. With technologies such as eye-tracking it might be possible to reveal which other landmarks speakers look at while or before they construct a referring expression.

Another possible explanation for the surprising outcome of our experiment is that our scenarios are too simple: they do not reflect the complexity of real-world visual scenes, and so the complex mechanisms we think are required for REG more generally are simply not required in these simple scenes. Rather than compute a reference that takes account of the context, the subjects in the iMAP Task perhaps recognise that the scenes are simple enough

to use referring expressions that are not carefully computed on the basis of context.

But this then raises a methodological issue. An assumption implicit in much recent work on evaluation in REG is that, by initially using simplistic domains and tasks, the in-principle capabilities of algorithms can be tested before scaling up to more complex real-world settings. The visual scenarios that are represented by the TUNA Corpus, the Drawer Corpus, and the GRE3D3 and GRE3D7 Corpora are very abstract and arguably quite unlike any real-world scenes where a speaker needs to construct a reference. For the work presented here, we attempted to consider more ‘realistic’ scenes involving speakers discussing larger numbers of objects in a distinct task; but even here, the scenario is still very simple with much fewer attributes to choose from than speaker are usually presented with when referring ‘in the wild’. But if this is the case, then what do we learn by developing algorithms that work in these simple scenarios?

We do not believe that the idea that human speakers deliberately build distinguishing descriptions in order to uniquely identify their intended referents should be abandoned: this seems to us a fundamentally important aspect of successful referential behaviour. But if we want to understand how it is that people do this, we should be wary of thinking we can learn about these processes by looking at how people refer in vastly simplified models of the real world. To move forward, we need to focus on the complexity of real-world reference scenarios.

7 Conclusions

Traditional REG algorithms are based on the aim of distinguishing the target referent from the other objects in its context. However, using a corpus of maptask dialogues, we found in earlier work that using features based on the same considerations as those underlying the traditional REG algorithms does not help in machine learning which attributes people use in a given situation. In this paper, we used two different methods of varying the size of the visual context that gets taken into account in computing the values for these features. We found that it is not possible to systematically determine an ideal context size using these methods, which seems to point to the conclusion that, for the speakers in our corpus, visual context was not an important consideration. Alternatively, even more fine-grained methods of determining the visual context than those we tried might be necessary, or the scenarios on the maps underlying our corpus are too simplistic to elicit real-world behaviour from the speakers. This points to the conclusion that it might be time for the field to move on to more complex visual scenes when researching content selection mechanisms for referring expression generation.

References

- Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.
- John M. Carroll. 1980. Naming and describing in social communication. *Language and Speech*, 23:309–322.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.
- Robert Dale and Nicholas Haddock. 1991. Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, pages 161–166, Berlin, Germany.
- 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 and Jette Viethen. 2010. Attribute-centric referring expression generation. In Emiel Krahmer and Mariët Theune, editors, *Empirical Methods in Natural Language Generation*, volume 5980 of *Lecture Notes in Computer Science*, pages 163–179. Springer.
- Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver BC, Canada.
- Claire Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 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 International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Sydney, Australia.
- Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 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*, pages 174–182, Athens, Greece.
- Martijn Goudbeek and Emiel Krahmer. 2010. Preferences versus adaptation during referring expression generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 55–59, Uppsala, Sweden.
- Barbara J. Grosz and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Markus Guhe. 2007. Marking theme and rheme in preverbal messages. In *Proceedings of the Seventh International Workshop on Computational Semantics*, pages 330–333, Tilburg, The Netherlands.
- Markus Guhe. 2009. Generating referring expressions with a cognitive model. In *Proceedings of the Workshop Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference*, Amsterdam, The Netherlands.
- Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications, Stanford CA, USA.
- Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.
- Max M. Louwerse, Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwyneth Lewis, Jie Wu, and Megan Zirnstein. 2007. Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1235–1240.
- Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–226.
- Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2006. Manual for the TUNA corpus: Referring expressions in two domains. Technical Report AUCS/TR0705, Computing Department, University of Aberdeen, UK.
- 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*, pages 63–70, Sydney, Australia.
- Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 59–67, Salt Fork OH, USA.
- Jette Viethen and Robert Dale. 2011. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the Workshop on Using Corpora in Natural Language Generation and Evaluation*, Edinburgh, UK.
- Jette Viethen, Robert Dale, and Markus Guhe. 2011. Generating subsequent reference in shared visual scenes: Computation vs. re-use. In *Proceeding the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK.
- Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco CA, USA.



ACL Member Portal

The Association for Computational Linguistics Member Portal

Portal Services

- Browse ACL Events
- Browse Non-ACL Events
- Browse the Entire Events Database
- Add an Event to the Database
- Browse the Participants List
- Advanced Search
- Contact Us

My Membership

- Manage My Journal Alerts
- Manage My Event Alerts
- View Membership Fees
- Pay My Membership and/or Donate
- Update My Profile

Other ACL Services

- Access the Computational Linguistics Journal
- Access Publishers' Discounts
- Visit the ACL Anthology

User login

Username: *

Password: *

- Create New Member Account
- Request New Password
- Request Username Reminder

Search

Search this site:

Home → 13th European Workshop on Natural Language Generation - submission deadline postponed to June 8th

13th European Workshop on Natural Language Generation - submission deadline postponed to June 8th

Submitted by Kristina I Stri... on 24 May 2011 - 10:46am

Tagged: natural language generation

Abbreviated Title: ENLG-11 - deadline postponed

Call for Papers

Submission Deadline: 8 Jun 2011

Event Dates: 28 Sep 2011 - 30 Sep 2011

City: Nancy

Country: France

Contact: Claire Gardent

Kristina Striegnitz

Contact Email: gardent@loria.fr

striegn@union.edu

Website: <http://tal.c.loria.fr/13th-European-Workshop-on-Natural.html>

***** Upon popular request, the deadline for submissions has been postponed

***** to June 8th, 2011

13th EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION (ENLG 2011)

<http://tal.c.loria.fr/13th-European-Workshop-on-Natural.html>

LAST CALL FOR PAPERS

ENLG 2011 will take place in Nancy, France, Sep. 28-30, 2011.

Submission deadline: June 8, 2011

Invited Speakers:

* Oliver Lemon, Heriot-Watt University, Edinburgh, UK

* Johanna Moore, University of Edinburgh, UK

* Jeff Orkin, MIT Media Labs, USA

The ENLG 2011 workshop continues a biennial series of workshops on natural language generation that has been running since 1987, providing a regular forum for presentation of research in this area. ENLG 2011 invites substantial, original, and unpublished submissions on all topics related to natural language generation. This includes research on "core" NLG issues as well as research in any area of NLP that produces language output, such as dialog systems, summarization, MT, etc. Topics of interest include but are not limited to:

- Generation of affect and emotion
- Personalization and personality of text
- Content and text planning
- Lexicalisation
- Referring expression generation
- Surface realization
- Generation for dialog systems
- Generation for embodied agents and robots
- Evaluation of NLG systems
- Text-to-text generation
- Generation for summarization
- Multimedia or multimodal generation
- Story-telling and narrative generation
- NLG for real-world applications e.g., Computer Assisted Language Learning, Authoring tools
- Psycholinguistics and NLG
- NLG in linguistically motivated frameworks
- Statistical processing for NLG
- Use of ontologies in NLG
- Generating controlled languages

ENLG 2011 will also include a special session on the Generation Challenges 2011. Generation Challenges is an umbrella event designed to bring together different shared-task evaluation efforts that involve the generation of natural language. In 2011 four shared-tasks on surface realisation, generating instructions in virtual environments, question generation, and the improvement of academic papers are being organized. Some of them will be presented at ENLG 2011. See the Generation Challenges website at <http://www.nltg.brighton.ac.uk/research/genchal11> for more information on how to participate in a shared task or propose a new shared task.

Requirements: Papers that have been or will be submitted to other meetings or publications must provide this information at submission time. If ENLG 2011 accepts a paper, authors must notify the program chairs, indicating which meeting they choose for presentation of their work. ENLG 2011 cannot accept for publication or presentation work that will be (or has been) published elsewhere.

Submission categories: ENLG invites the submission of long and short papers:

- Long papers are most appropriate for presenting substantial research results and must not exceed eight (8) pages, excluding references (accepted long papers will be presented orally);

- Short papers are more appropriate for presenting an ongoing research effort and must not exceed four (4) pages, excluding references (these will be presented as posters during the poster session).

Paper Submission: Submissions should be uploaded to the ENLG 2011 submission site, which will be made available. Submissions must conform to the official ACL-HLT 2011 style guidelines (see <http://www.acl2011.org/call.shtml#submission>), which are contained in the style files, and they must be electronic in PDF. **Reviewing will be blind, so you should avoid identifying the authors within the paper. Please note that different guidelines and procedures apply to papers for the shared tasks. Please refer to the organizers of the different challenges.**

If you have any questions, please feel free to contact the workshop organizers:

Claire Gardent
CNRS/Loria Nancy, France
gardent@loria.fr

Kristina Striegnitz
Union College, Schenectady, NY, USA
striegn@union.edu

Important dates:

- June 8, 2011 Deadline for paper submission
- July 15, 2011 Notification of acceptance of papers
- Sep 1, 2011 Camera-ready copies due
- Sep 28-30, 2011 ENLG 2011

The following researchers have agreed to be members of the ENLG 2011 Program Committee.

- John Bateman, University Bremen, Germany
- Anja Belz, University of Brighton, UK
- Bernd Bohnet, University Stuttgart, Germany
- Stephan Busemann, DFKI, Germany
- Charles Callaway, University of Edinburgh, UK
- Christian Chiarcos, University of Potsdam, Germany
- Norman Creaney, University of Ulster, Ireland
- Robert Dale, Macquarie University, Australia
- Kees van Deemter, University of Aberdeen, Scotland
- Seniz Demir, University of Delaware, USA
- Alexandre Denis, CNRS/LORIA Nancy, France
- David DeVault, USC Institute for Creative Technologies, USA
- Barbara Di Eugenio, University of Illinois, USA
- Roger Evans, University of Brighton, UK
- Leo Ferres, University of Concepcion, Chile
- Jennifer Foster, Dublin University, Ireland
- Mary Ellen Foster, Heriot Watt University, Edinburgh, Scotland
- Claire Gardent, CNRS/LORIA, France
- Albert Gatt, University of Malta, Malta
- Josef van Genabith, Dublin City University, Ireland

- Pablo Gervas, Universidad Complutense de Madrid, Spain
- Markus Guhe, University of Edinburgh, UK
- John Kelleher, Dublin Institute of Technology, Ireland
- Alistair Knott, University of Otago, New Zealand
- Alexander Koller, University of Saarbrücken, Germany
- Stefan Kopp, University of Bielefeld, Germany
- Eric Kow, University of Brighton, UK
- Emiel Krahmer, Tilburg University, The Netherlands
- Geert-Jan Kruijff, DFKI, Germany
- Ivana Kruijff-Korbayova, DFKI, Germany
- Oliver Lemon, Heriot Watt University, Edinburgh, Scotland
- James Lester, North Carolina State University, USA
- Keith van der Linden, Calvin College, USA
- François Mairesse, University of Cambridge, UK
- Kathleen McCoy, University of Delaware, USA
- David McDonald, SIFT, Inc., USA
- Chris Mellish, University of Aberdeen, Scotland
- Jon Oberlander, University of Edinburgh, Scotland
- Cécile Paris, CSIRO ICT Centre, Australia
- Paul Piwek, The Open University, UK
- Richard Power, The Open University, UK
- Ehud Reiter, University of Aberdeen and Data2Text Ltd, Scotland
- Donia Scott, University of Sussex, UK
- Advait Siddharthan, University of Aberdeen, Scotland
- Ielka van der Sluis, Trinity College Dublin, Ireland
- Yaji Sripada, University of Aberdeen, Scotland
- Manfred Stede, University of Potsdam, Germany
- Amanda Stent, AT&T Labs - Research, USA
- Matthew Stone, Rutgers, USA
- Kristina Striegnitz, Union College, Schenectady, USA
- Michael Strube, EML Research, Germany
- Mariët Theune, University of Twente, The Netherlands
- Takenobu Tokugana, Tokyo Institute of Technology, Japan
- Jette Viethen, Macquarie University, Australia
- Carl Vogel, Trinity College Dublin, Ireland
- Michael White, Ohio State University, USA
- Sandra Williams, the Open University, UK
- Tie-Jun Zhao, Harbin Institute of Technology, China
- Michael Zock, CNRS/LIF Université de la Méditerranée Aix-Marseille II, France

ENLG 2011 is endorsed by the ACL Special Interest Group on Generation (SIGGEN).

ENLG 2011

**The 13th European
Workshop on Natural
Language Generation**

Proceedings

28 – 30 September 2011

ENLG 2011 is sponsored by



Nancy-Université



©2011 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Introduction

It is with great pleasure that we present the current volume of papers accepted for presentation at the 13th European Workshop on Natural Language Generation (ENLG 2011), which will be held from September 28th to 30th, 2011 at Loria in Nancy, France.

The ENLG 2009 workshop continued a biennial series of workshops on natural language generation that has been running since 1987 and alternates with INLG, the International Conference on Natural Language Generation. Previous European workshops have been held at Royaumont, Edinburgh, Judenstein, Pisa, Leiden, Duisburg, Toulouse, Budapest, Aberdeen, Dagstuhl and Athens. Together with INLG, the ENLG workshop is the main regular forum for presenting and discussing research in Natural Language Generation.

ENLG 2011 invited submissions on all topics related to natural language generation. We received 41 submissions of long and short papers from all over the world. Of these 13 long papers and 12 short papers were accepted for presentation. The long papers will be presented orally, and the short papers as posters.

In addition, ENLG 2011 hosts Generation Challenges 2011. This year, three shared task evaluation competitions were organized under the umbrella of Generation Challenges 2011: the Surface Realisation Challenge (Belz, Hogan, White, and Stent), the Challenge on Generating Instructions in Virtual Environments (Striegnitz, Denis, Gargett, Garoufi, Koller, and Theune) and the Helping Our Own Challenge (Dale and Kilgariff).

The first part of this volume contains the 25 research papers that will be presented at ENLG 2011. The second part is devoted to the Generation Challenges 2011 session. It contains overview reports on the active and planned challenges and system descriptions of all participating teams.

We are indebted to the authors and to the members of our program committee whose hard work contributed to making this a collection of high quality research papers. We are also delighted that Oliver Lemon, Johanna Moore and Jeff Orkin agreed to give invited talks at ENLG 2011. And last but not least, many thanks go to the local organisation team, Nicolas Alcaraz, Anne Lise Charbonnier, Alexandre Denis, Alejandra Lorenzo, Shashi Narayan and Laura Perez-Beltrachini for handling the preparation of the meeting.

Claire Gardent and Kristina Striegnitz
Program co-Chairs for ENLG 2011

Program Co-Chairs:

Claire Gardent, CNRS/Loria, Nancy, France

Kristina Striegnitz, Union College, USA

Local Organizers:

Nicolas Alcaraz, INRIA Nancy, France

Anne-Lise Charbonnier, INRIA Nancy, France

Alexandre Denis, CNRS/LORIA, Nancy, France

Alejandra Lorenzo, INRIA/LORIA, Nancy, France

Shashi Narayan, U. Nancy/LORIA, Nancy, France

Laura Perez-Beltrachini, U. Nancy/LORIA, Nancy, France

Program Committee:

John Bateman, University Bremen, Germany

Anja Belz, University of Brighton, UK

Bernd Bohnet, University Stuttgart, Germany

Stephan Busemann, DFKI, Germany

Christian Chiarcos, University of Potsdam, Germany

Norman Creaney, University of Ulster, Ireland

Robert Dale, Macquarie University, Australia

Kees van Deemter, University of Aberdeen, Scotland

Seniz Demir, Tubitak-Bilgem, Turkey

Alexandre Denis, CNRS/LORIA Nancy, France

David DeVault, USC Institute for Creative Technologies, USA

Barbara Di Eugenio, University of Illinois, USA

Roger Evans, University of Brighton, UK

Leo Ferres, University of Concepcion, Chile

Jennifer Foster, Dublin University, Ireland

Albert Gatt, University of Malta, Malta

Josef van Genabith, Dublin City University, Ireland

Pablo Gervas, Universidad Complutense de Madrid, Spain

Markus Guhe, University of Edinburgh, UK

John Kelleher, Dublin Institute of Technology, Ireland

Alistair Knott, University of Otago, New Zealand

Alexander Koller, University of Saarbrcken, Germany

Stefan Kopp, University of Bielefeld, Germany

Eric Kow, University of Brighton, UK

Emiel Kraemer, Tilburg University, The Netherlands

Geert-Jan Kruijff, DFKI, Germany

Ivana Kruijff-Korbayova, DFKI, Germany

Oliver Lemon, Heriot Watt University, Edinburgh, Scotland

James Lester, North Carolina State University, USA
Keith van der Linden, Calvin College, USA
François Mairesse, University of Cambridge, UK
Kathleen McCoy, University of Delaware, USA
David McDonald, SIFT, Inc., USA
Chris Mellish, University of Aberdeen, Scotland
Jon Oberlander, University of Edinburgh, Scotland
Cécile Paris, CSIRO ICT Centre, Australia
Paul Piwek, The Open University, UK
Richard Power, The Open University, UK
Ehud Reiter, University of Aberdeen and Data2Text Ltd, Scotland
Donia Scott, University of Sussex, UK
Advaith Siddharthan, University of Aberdeen, Scotland
Ielka van der Sluis, Trinity College Dublin, Ireland
Yaji Sripada, University of Aberdeen, Scotland
Manfred Stede, University of Potsdam, Germany
Amanda Stent, AT&T Labs Research, USA
Matthew Stone, Rutgers, USA
Michael Strube, EML Research, Germany
Mariët Theune, University of Twente, The Netherlands
Takenobu Tokugana, Tokyo Institute of Technology, Japan
Jette Viethen, Macquarie University, Australia
Carl Vogel, Trinity College Dublin, Ireland
Michael White, Ohio State University, USA
Sandra Williams, the Open University, UK
Tie-Jun Zhao, Harbin Institute of Technology, China
Michael Zock, CNRS/LIF Université de la Méditerranée Aix-Marseille II, France

Invited Speakers:

Oliver Lemon, Heriot Watt University, Scotland
Johanna Moore, University of Edinburgh, Scotland
Jeff Orkin, MIT Media Lab, USA

Table of Contents

Oral Presentations

<i>Talkin' bout a revolution (statistically speaking)</i> (Invited Talk) Oliver Lemon	1
<i>Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies</i> Advait Siddharthan	2
<i>Generating Affective Natural Language for Parents of Neonatal Infants</i> Saad Mahamood and Ehud Reiter	12
<i>What is in a text and what does it do: Qualitative Evaluations of an NLG system – the BT-Nurse – using content analysis and discourse analysis</i> Rahul Sambaraju, Ehud Reiter, Robert Logie, Andy Mckinlay, Chris McVittie, Albert Gatt and Cindy Sykes	22
<i>Evaluating Saliency Metrics for the Context-Adequate Realization of Discourse Referents</i> Christian Chiarcos	32
<i>The Impact of Visual Context on the Content of Referring Expressions</i> Henriette Viethen, Robert Dale and Markus Guhe	44
<i>A Cross-Linguistic Study on the Production of Multimodal Referring Expressions in Dialogue</i> Ielka Van Der Sluis and Saturnino Luz	53
<i>Two Approaches for Generating Size Modifiers</i> Margaret Mitchell, Kees Van Deemter and Ehud Reiter	63
<i>Using Online Games to Capture, Generate, and Understand Natural Language</i> (Invited Talk) Jeff Orkin	71
<i>Content selection from an ontology-based knowledge base for the generation of football summaries</i> Nadjet Bouayad-Agha, Gerard Casamayor and Leo Wanner	72
<i>Deriving rhetorical relationships from semantic content</i> Richard Power	82
<i>If it may have happened before, it happened, but not necessarily before</i> Albert Gatt and François Portet	91
<i>Adaptive Information Presentation for Spoken Dialogue Systems: Evaluation with real users</i> Verena Rieser, Simon Keizer, Oliver Lemon and Xingkun Liu	102

<i>Combining Hierarchical Reinforcement Learning and Bayesian Networks for Natural Language Generation in Situated Dialogue</i>	
Nina Dethlefs and Heriberto Cuayáhuitl	110
<i>Combining symbolic and corpus-based approaches for the generation of successful referring expressions</i>	
Konstantina Garoufi and Alexander Koller	121
<i>Language Generation for Spoken Dialogue Systems</i> (Invited Talk)	
Johanna D. Moore	132
Poster Presentations	
<i>Adapting SimpleNLG to German</i>	
Marcel Bollmann	133
<i>EasyText: an Operational NLG System</i>	
Laurence Danlos, Frédéric Meunier and Vanessa Combet	139
<i>Towards Generating Text from Discourse Representation Structures</i>	
Valerio Basile and Johan Bos	145
<i>A Policy-Based Approach to Context Dependent Natural Language Generation</i>	
Thomas Bouttat, Edoardo Pignotti, Chris Mellish and Peter Edwards	151
<i>Levels of organisation in ontology verbalisation</i>	
Sandra Williams, Allan Third and Richard Power	158
<i>Using semantic roles to improve summaries</i>	
Diana Trandabâţ	164
<i>Building a Generator for Italian Sign Language</i>	
Alessandro Mazzei	170
<i>Investigation into Human Preference between Common and Unambiguous Lexical Substitutions</i>	
Andrew Walker, Advaith Siddharthan and Andrew Starkey	176
<i>Production of Demonstratives in Dutch, English and Portuguese Dialogues</i>	
Saturnino Luz and Ielka Van Der Sluis	181
<i>Generation of Formal and Informal Sentences</i>	
Fadi Abu Sheikha and Diana Inkpen	187
<i>Glue Rules for Robust Chart Realization</i>	
Michael White	194
<i>Detecting Interesting Event Sequences for Sports Reporting</i>	
François Lareau, Mark Dras and Robert Dale	200

Generation Challenges 2011

<i>Generation Challenges 2011 Preface</i>	
Anja Belz, Albert Gatt, Alexander Koller and Kristina Striegnitz	206
<i>The GRUVE Challenge: Generating Routes under Uncertainty in Virtual Environments</i>	
Srini Janarthanam and Oliver Lemon	208
<i>A Proposal for a Spanish Surface Realization Shared Task</i>	
Pablo Gervás and Miguel Ballesteros	212
<i>The First Surface Realisation Shared Task: Overview and Evaluation Results</i>	
Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan and Amanda Stent	217
<i>DCU at Generation Challenges 2011 Surface Realisation Track</i>	
Yuqing Guo, Deirdre Hogan and Josef van Genabith	227
<i>ATT-0: Submission to Generation Challenges 2011 Surface Realization Shared Task</i>	
Amanda Stent	230
<i><StuMaBa >: From Deep Representation to Surface</i>	
Bernd Bohnet, Simon Mille, Benoît Favre and Leo Wanner	232
<i>The OSU System for Surface Realization at Generation Challenges 2011</i>	
Rajakrishnan Rajkumar, Dominic Espinosa and Michael White	236
<i>UCM Submission to the Surface Realization Challenge</i>	
Pablo Gervás	239
<i>Helping Our Own: The HOO 2011 Pilot Shared Task</i>	
Robert Dale and Adam Kilgarriff	242
<i>May I check the English of your paper!!!</i>	
Pinaki Bhaskar, Aniruddha Ghosh, Santanu Pal and Sivaji Bandyopadhyay	250
<i>Handling Outlandish Occurrences: Using Rules and Lexicons for Correcting NLP Articles</i>	
Elitza Ivanova, Delphine Bernhard and Cyril Grouin	254
<i>NUS at the HOO 2011 Pilot Shared Task</i>	
Daniel Dahlmeier, Hwee Tou Ng and Thanh Phu Tran	257
<i>Helping Our Own 2011: UKP Lab System Description</i>	
Torsten Zesch	260
<i>University of Illinois System in HOO Text Correction Shared Task</i>	
Alla Rozovskaya, Mark Sammons, Joshua Gioja and Dan Roth	263
<i>Data-Driven Correction of FunctionWords in Non-Native English</i>	
Adriane Boyd and Detmar Meurers	267

<i>Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5)</i>	
Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller and Mariët Theune	270
<i>Direction giving: an attempt to increase user engagement</i>	
Bob Duncan and Kees van Deemter	280
<i>The Bremen System for the GIVE-2.5 Challenge</i>	
Nina Dethlefs	284
<i>The GIVE-2.5 C Generation System</i>	
David Nicolás Racca, Luciana Benotti and Pablo Duboue	290
<i>CL system: Giving instructions by corpus based selection</i>	
Luciana Benotti and Alexandre Denis	296
<i>The Loria Instruction Generation System L in GIVE 2.5</i>	
Alexandre Denis	302
<i>The Potsdam NLG systems at the GIVE-2.5 Challenge</i>	
Konstantina Garoufi and Alexander Koller	307
<i>The Thumbs Up! Twente system for GIVE 2.5</i>	
Saskia Akkersdijk, Marin Langenbach, Frieder Loch and Mariët Theune	312
<i>Question Generation Shared Task and Evaluation Challenge – Status Report</i>	
Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev and Cristian Moldovan	318

Conference Program

Wednesday, September 28, 2011

8:30–9:15 Registration

9:15–9:30 Opening Remarks

9:30–10:30 Invited Talk

Talkin' bout a revolution (statistically speaking)

Oliver Lemon

10:30–11:00 Coffee Break

ENLG Talks: Tailoring text to users

11:00–11:30 *Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies*

Advaith Siddharthan

11:30–12:00 *Generating Affective Natural Language for Parents of Neonatal Infants*

Saad Mahamood and Ehud Reiter

12:00–12:30 *What is in a text and what does it do: Qualitative Evaluations of an NLG system – the BT-Nurse – using content analysis and discourse analysis*

Rahul Sambaraju, Ehud Reiter, Robert Logie, Andy Mckinlay, Chris McVittie, Albert Gatt and Cindy Sykes

12:30–1:30 Lunch

ENLG Talks: Referring expression generation

1:30–2:00 *Evaluating Saliency Metrics for the Context-Adequate Realization of Discourse Referrers*

Christian Chiarcos

2:00–2:30 *The Impact of Visual Context on the Content of Referring Expressions*

Henriette Viethen, Robert Dale and Markus Guhe

2:30–3:00 Coffee Break

Wednesday, September 28, 2011 (continued)

ENLG Talks: Referring expression generation

- 3:00–3:30 *A Cross-Linguistic Study on the Production of Multimodal Referring Expressions in Dialogue*
Ielka Van Der Sluis and Saturnino Luz
- 3:30–4:00 *Two Approaches for Generating Size Modifiers*
Margaret Mitchell, Kees Van Deemter and Ehud Reiter
- 1:30-2:30 Birds-of-a-Feather sessions / SimpleNLG User Group Meeting
- 6:30-8:00 Bowling and Drinks

Thursday, September 29, 2011

- 9:30–10:30 Invited Talk
Using Online Games to Capture, Generate, and Understand Natural Language
Jeff Orkin
- 10:30–11:00 Coffee Break
- ENLG Talks: Knowledge representation and NLG — expressing semantic, rhetorical and temporal relations**
- 11:00–11:30 *Content selection from an ontology-based knowledge base for the generation of football summaries*
Nadjet Bouayad-Agha, Gerard Casamayor and Leo Wanner
- 11:30–12:00 *Deriving rhetorical relationships from semantic content*
Richard Power
- 12:00–12:30 *If it may have happened before, it happened, but not necessarily before*
Albert Gatt and François Portet
- 12:30–1:30 Lunch

Thursday, September 29, 2011 (continued)

Generation Challenges Talks and Posters

- 1:30–1:40 Generation Challenges Introductory Remarks
- 1:40–2:55 Result Presentation: Surface Realization (SR), Helping Our Own (HOO), Generating Instructions in Virtual Environments (GIVE)
- 2:55–3:40 Presentation of Planned Tasks: SR Spanish, Question Generation 2nd Edition, Generating Route Instructions under Uncertainty in Virtual Environments (GRUVE)
- 3:40–5:00 Generation Challenges Poster Session (with coffee)
- 6:00–7:00 Guided Tour in the Old Town
- 7:30 Banquet

Friday, September 30, 2011

ENLG Talks: Optimizing task success in interactive systems

- 9:30–10:00 *Adaptive Information Presentation for Spoken Dialogue Systems: Evaluation with real users*
Verena Rieser, Simon Keizer, Oliver Lemon and Xingkun Liu
- 10:00–10:30 *Combining Hierarchical Reinforcement Learning and Bayesian Networks for Natural Language Generation in Situated Dialogue*
Nina Dethlefs and Heriberto Cuayáhuitl
- 10:30–11:00 *Combining symbolic and corpus-based approaches for the generation of successful referring expressions*
Konstantina Garoufi and Alexander Koller
- 11:00–12:30 ENLG Posters (see below for a detailed list)
- 12:30–1:30 Lunch

Friday, September 30, 2011 (continued)

- 1:30–2:30 Invited Talk
- Language Generation for Spoken Dialogue Systems*
Johanna D. Moore
- 2:30–3:00 Coffee Break
- 3:00–5:00 Generation Challenges Working Groups
- 5:00–5:30 Reports from the Working Groups and Closing Remarks

ENLG Poster Presentations

Adapting SimpleNLG to German
Marcel Bollmann

EasyText: an Operational NLG System
Laurence Danlos, Frédéric Meunier and Vanessa Combet

Towards Generating Text from Discourse Representation Structures
Valerio Basile and Johan Bos

A Policy-Based Approach to Context Dependent Natural Language Generation
Thomas Bouttaz, Edoardo Pignotti, Chris Mellish and Peter Edwards

Levels of organisation in ontology verbalisation
Sandra Williams, Allan Third and Richard Power

Using semantic roles to improve summaries
Diana Trandabăt

Building a Generator for Italian Sign Language
Alessandro Mazzei

Investigation into Human Preference between Common and Unambiguous Lexical Substitutions
Andrew Walker, Advait Siddharthan and Andrew Starkey

Friday, September 30, 2011 (continued)

Production of Demonstratives in Dutch, English and Portuguese Dialogues
Saturnino Luz and Ielka Van Der Sluis

Generation of Formal and Informal Sentences
Fadi Abu Sheikha and Diana Inkpen

Glue Rules for Robust Chart Realization
Michael White

Detecting Interesting Event Sequences for Sports Reporting
François Lareau, Mark Dras and Robert Dale

