Evaluating algorithms for the Generation of Referring Expressions using a balanced corpus

Albert Gatt and **Ielka van der Sluis** and **Kees van Deemter**
Department of Computing Science
University of Aberdeen
{agatt,ivdsluis,kvdeemte}@csd.abdn.ac.uk

Abstract

Despite being the focus of intensive research, *evaluation* of algorithms that generate referring expressions is still in its infancy. We describe a corpus-based evaluation methodology, applied to a number of classic algorithms in this area. The methodology focuses on *balance* and *semantic transparency* to enable comparison of human and algorithmic output. Although the Incremental Algorithm emerges as the best match, we found that its dependency on manually-set parameters makes its performance difficult to predict.

1 Introduction

The current state of the art in the Generation of Referring Expressions (GRE) is dominated by versions of the Incremental Algorithm (IA) of Dale and Reiter (1995). Focusing on the generation of "first-mention" definite descriptions, Dale and Reiter compared the IA to a number of its predecessors, including a Full Brevity (FB) algorithm, which generates descriptions of minimal length, and a Greedy algorithm (GR), which approximates Full Brevity (Dale, 1989). In doing so, the authors focused on Content Determination (CD, which is the purely semantic part of GRE), and on a description's ability to *identify* a referent for a hearer. Under this problem definition, GRE algorithms take as input a Knowledge Base (KB), which lists domain entities and their properties (often represented as attribute-value pairs), together with a set of intended referents, R. The output of CD is a distinguishing description of R, that is, a logical form which distinguishes this set from its distractors.

Dale and Reiter argued that the IA was a superior model, and predicted that it would be the better match to human referential behaviour.[1] This was due in part to the way the IA searches for a distinguishing description by performing gradient descent along a predetermined list of domain attributes, called the *preference order*, whose ranking reflects general or domain-specific preferences (see §4.1).

The Incremental Algorithm has served as a starting point for later models (Horacek, 1997; Kelleher and Kruijff, 2006), and has also served as a yardstick against which to compare other approaches (Gardent, 2002; Jordan and Walker, 2005). Despite its influence, few empirical evaluations have focused on the IA. Evaluation is even more desirable given the dependency of the algorithm on a preference order, which can radically change its behaviour, so that in a domain with n attributes, there are in principle $n!$ different algorithms.

This paper is concerned with applying a corpus-based methodology to evaluate content determination for GRE, comparing the three classic algorithms that formed the basis for Dale and Reiter's (1995) contribution, adapted to also deal with pluralities and gradable properties.

1.1 Requirements for GRE evaluation

One of the problems with evaluating GRE is that it interfaces with several other sub-tasks of NLG including, among others, realisation and discourse coherence (especially where anaphoric reference is concerned). On the other hand, a large amount of work in the area has focused on the semantic heart of the problem. Given *identification* as the over-

[1] Dale and Reiter also observed that IA is computationally more efficient than its competitors, although GR has only polynomial complexity. Consistent with subsequent research, we shall be de-emphasising complexity issues here.

arching goal of such algorithms, a crucial question concerns the extent to which their choice of content from the available attributes for a referent matches that produced by a speaker in a comparable situation. This is the main focus of this paper, whose evaluation methodology therefore targets content determination, abstracting away from issues of lexical choice and realisation. A corpus-based evaluation of a content determination GRE algorithm requires a resource that satisfies the following desiderata.

Semantic transparency: The human 'gold standard' descriptions in the corpus need to be paired with a domain representation so that, as far as possible, an algorithm is exposed to the same domain as an author. To evaluate content determination, descriptions need to be semantically annotated, abstacting away from variations in syntax and lexicalisation. For example, *the right-facing sofa* and *the settee which is oriented towards the right* are, from the point of view of a content determination procedure, semantically equivalent.

Pragmatic transparency: Ideally, the communicative intention underlying corpus descriptions should match those for which an algorithm was designed. If an algorithm is primarily aimed at identification, then human gold-standards should, as far as possible, be restricted to this intention.

Balance: To assess the extent to which an algorithm matches human performance, the corpus should contain an equal number of instances where each attribute is required. Only in this way would the claim that *algorithm* X *matches humans on content y% of the time* be reliable[2].

These desiderata suggest that the way forward in evaluation in this area is to design controlled studies for corpus construction. The rest of this paper describes the construction of such a corpus, and the results of an evaluation that addressed the differences between IA and its predecessors against human descriptions in domains of varying complexity, containing both singular and plural descriptions. The study also aimed to contribute to a growing debate in the NLG community, on the evaluation of NLG systems, arguing in favour of the careful construction of *balanced* and *transparent* corpora to serve as resources for NLG.

[2]For example, the IA overspecifies descriptions by selecting attributes not strictly required for identification, because of its preference order. A claim that this feature improves performance implies that the relative priority of attributes is important. To be reliable, such a claim would have to be made against a corpus in which 'preferred' and 'dispreferred' attributes were required the same number of times.

2 Related work

We are aware of three studies on GRE evaluation, all of which compare the IA to some alternative models. Two of these (Jordan and Walker, 2005; Gupta and Stent, 2005) used the COCONUT dialogue corpus. The third (Viethen and Dale, 2006) used a small corpus collected in a monologue setting. These studies meet the transparency requirements to different degrees. Though COCONUT dialogues were elicited against a well-defined domain, Jordan (2000) has emphasised that reference, in COCONUT, was often intended to satisfy intentions over and above identification. Gupta and Stent used an evaluation metric that included aspects of the syntactic structure of descriptions (specifically, modifier placement), thus arguably obscuring the role of content determination (CD).

Our approach is closest in spirit to that of Viethen and Dale, who elicited descriptions from people in a setting where identification was the sole communicative aim. However, in the case of the IA, the authors averaged over 24 different preference orders, potentially averaging over 24 very different incarnations of the algorithm and masking the impact of any one order. Similarly, neither Jordan/Walker nor Gupta/Stent are explicit about the determination of the preference order for the IA in their studies. No obvious attempts were made to make sure that the corpora in question were semantically balanced.

One question that these studies raise relates to how human-authored and automatically generated descriptions should be compared. For instance, both Jordan/Walker and Viethen/Dale use a measure of recall. This indicates the coverage of an algorithm in relation to a corpus, but does not measure the *degree* of similarity between a description generated by an algorithm and a description in the corpus, punishing all mismatches with equal severity.

3 The TUNA corpus

We built a corpus consisting of ca. 1800 descriptions, collected through a controlled experiment run over the web for three months. Half of this corpus contains descriptions of real photographs of people; the other half contains descriptions of artificially constructed pictures of household items. In this paper, the main focus is on the 'furniture' subcorpus,

TYPE	COLOUR	ORIENTATION	SIZE
chair	blue	forward	large
sofa	red	backward	small
desk	green	leftward	
fan	grey	rightward	

Table 1: Non-numeric attributes in the domains

which represents the simpler of the two domains, consisting of digitally constructed pictures of objects with well-defined properties. Therefore, it provides a good test case for the algorithms evaluated, since it allows us to probe into a number of issues that arise even with straightforwardly describable objects. The 'people' sub-corpus is more complex, since the objects are real photographs and afford an author with many descriptive alternatives. We explicitly compare the results of the present evaluation with a similar study on the 'people' sub-corpus, in §5.

3.1 Materials, design and procedure

The furniture sub-corpus consists of 900 descriptions from 45 native or fluent speakers of English. Participants described objects in 20 trials, each corresponding to a domain where there were one or two clearly marked target referents (the *target set*) and six distractor objects, placed in a 3 (row) $\times$ 5 (column) grid. Pictures of the objects represented combinations of values of the four attributes shown in the top panel of Table 1. In a pilot study involving 19 participants, we found that instances in which descriptions used semantic content beyond that indicated in the Table were extremely rare with these simple objects. In each trial, the horizontal and vertical position of the objects is represented using two numeric-valued attributes, X-DIM (row) and Y-DIM (column). Their value was randomly determined with every fresh trial. Approximately half the corpus descriptions include locative expressions[3]. We will refer to this as the +LOC dataset, containing 412 descriptions from 26 authors. The other half, the −LOC dataset (444 descriptions; 27 authors), consists of descriptions using only COLOUR, SIZE and ORIENTATION, apart from TYPE.

Participants were exposed to the 20 trials in randomised order; in each case, they typed a description for the target set. They were told that they would be interacting with a language-understanding program which would remove the referents from the domain, based on their description. Identification was emphasised as the primary goal of descriptions. Each time a participant submitted a description, one or two objects were automatically removed from the domain by a function which had been preset to remove the wrong objects on approximately one-fourth of the trials. This was intended to make the interaction seem more natural. We discuss an evaluation of this methodology in §3.3.

The trials in the experiment were balanced. For each possible combination of the attributes in Table 1, there was an equal number of domains in which an identifying description of the target(s) required the use of those attributes. We refer to this as the *minimal description* (MD) of the target set. For example, there was a domain in which a target could be minimally distinguished by using COLOUR and SIZE. TYPE was never included in the minimal description, leaving 7 possible attribute combinations. The experiment manipulated one within-subjects variable, Cardinality/Similarity (3 levels):

Singular (SG): 7 domains contained a single referent

Plural/Similar (PS): 6 domains had two referents, which had identical values on the MD attributes. For example, both targets might be blue in a domain where the minimally distinguishing description consisted of COLOUR.

Plural/Dissimilar (PD): In the remaining 7 Plural trials, the targets had different values of the minimally distinguishing attributes.

Plural referents were taken into account because plurality is pervasive in NL discourse. The literature (e.g. Gardent (2002)) suggests that they can be treated adequately by minor variations of the classic GRE algorithms (as long as the descriptions in question refer distributively, cf. Stone (2000)), which is something we considered worth testing.

3.2 Corpus annotation

The XML annotation scheme (van der Sluis et al., 2006) pairs each corpus description with a representation of the domain in which it was produced. The domain representation, exemplified in Figure 1(a)), indicates which entities are target referents or distractors, and what combination of the attributes and values in Table 1 they have, as well as their numeric X-DIM and Y-DIM values (row and column numbers).

[3]This was manipulated as a second, between-subjects factor. Participants were randomly placed in groups which varied in whether they could use location or not, and in whether the communicative situation was fault-critical or not. For more details, we refer to van Deemter *et al.* (2006).

```
<ENTITY type=target'>
<ATTRIBUTE name='orientation' value='right' />
<ATTRIBUTE name='type' value='sofa' />
<ATTRIBUTE name='size' value='large' />
...
</ENTITY>
<ENTITY type='target'>
<ATTRIBUTE name='colour' value='red' />
<ATTRIBUTE name='type' value='desk' />
<ATTRIBUTE name='size' value='small' />
...
</ENTITY>
```

(a) Fragment of a domain

```
<DESCRIPTION num='plural'>
<DESCRIPTION num='singular'>
<ATTRIBUTE name='size' value='large'>large</ATTRIBUTE>
<ATTRIBUTE name='type' value='sofa'>settee</ATTRIBUTE>
<ATTRIBUTE name='orientation' value='right'>
at oblique angle</ATTRIBUTE>
</DESCRIPTION>
and
<DESCRIPTION num='singular'>
<ATTRIBUTE name='size' value='small'>small</ATTRIBUTE>
<ATTRIBUTE name='type' value='desk'>desk</ATTRIBUTE>
</DESCRIPTION>
</DESCRIPTION>
```

(b) 'large settee at oblique angle and small desk'

Figure 1: Corpus annotation examples

Figure 1(b) shows the annotation of a plural description. `ATTRIBUTE` tags enclose segments of a description corresponding to properties, with `name` and `value` attributes which constitute a semantic representation compatible with the domain, abstracting away from lexical variation. For example, in Figure 1(b), the expression *at an oblique angle* is tagged as ORIENTATION, with the value *rightward*. If a part of a description could not be resolved against the domain representation, it was enclosed in an `ATTRIBUTE` tag with the value `other` for `name`. Consistent with our pilot study, this was only necessary in 39 descriptions (3.2%).

The `DESCRIPTION` tag in Figure 1(b) indicates the logical form of a description. Thus, Figure 1(b) is a `plural` description enclosing two `singular` ones. Correspondingly, the logical form of each embedded description is a conjunction of attributes, while the two sibling descriptions are disjoined:

$$(\textit{large} \wedge \textit{sofa} \wedge \textit{right}) \vee (\textit{small} \wedge \textit{desk})\,.$$

3.3 Annotator reliability and experimental validity

The reliability of the annotation scheme was evaluated in a study involving two independent annotators (hereafter A and B), both postgraduate students with an interest in NLG, who used the same annotation manual (van der Sluis et al., 2006). They were given a stratified random sample of 270 descriptions, 2 from each Cardinality/Similarity condition, from each author in the corpus. To estimate inter-annotator agreement, we compared their annotations against the consensus labelling made by the present authors, using a version of the Dice coefficient. Let D_1 and D_2 be two descriptions, and $att(D)$ be the attributes in any description D. The coefficient, which ranges between 0 (no agreement) and 1 (perfect agreement) is calculated as in (1). Because descriptions could contain more than one instance of an attribute (e.g. Figure 1(b) contains two instances of SIZE), the sets of attributes for this comparison were represented as multisets.

$$dice(D_1, D_2) = \frac{2 \times |att(D_1) \cap att(D_2)|}{|att(D_1)| + |att(D_2)|} \qquad (1)$$

In the present context, Dice is more appropriate than agreement measures (such as the κ statistic) which rely on predefined categories in which discrete events can be classified. The 'events' in the corpus are NL expressions, each of which is 'classified' in several ways (depending on how many attributes a description expresses), and it was up to an annotator's judgment, given the instructions, to select those segments and mark them up.

Both annotators showed a high mean agreement with the authors, as indicated by their mean and modal (most frequent) scores (A:: mean $= 0.93$, mode $= 1$ (74.4%); B: mean $= 0.92$; mode $= 1$ (73%)). They also evinced substantial agreement among themselves (mean $= 0.89$, mode $= 1$ (71.1%)). These results suggest that the annotation scheme used is replicable to a high degree, and that independent annotators are likely to produce very similar semantic markup.

In the evaluation study reported below, we use the same measure to compare algorithm and human output, because an optimally informative comparison

should take into account the number of attributes that an algorithm omits in relation to the human gold standard, and the number of attributes that it includes. We also evaluated the validity of the experimental setup. Since communicating with a machine may have biased participants, they were asked, during a debriefing phase, to assess the performance of their virtual interlocutor, by indicating agreement to the statement *The system performed well on this task*. Of the 5 response categories, ranging from 1 *strongly disagree* to 5 (*strongly agree*), 34 individuals selected *agree* or *strongly agree* while no one selected *strongly disagree*. The mean score was 3.9.

4 Evaluating the algorithms

The three algorithms mentioned in §1 can be characterised as search problems (Bohnet and Dale, 2005) which differ primarily in the way they structure a search space populated by KB properties:

Full Brevity (FB): Finds the smallest distinguishing combination of properties.
Greedy (GR): Adds properties to a description, always selecting the property with the greatest discriminatory power.
Incremental (IA): Performs gradient descent along a predefined list of properties. Like GR, IA incrementally adds properties to a description until it is distinguishing.

The evaluation was carried out separately for the −LOC and +LOC datasets introduced in §3.1. Algorithms were compared to a random baseline (RAND) which selected a property randomly, and added it to a description if it removed distractors and was true of the referents. In the −LOC dataset, only GR and IA were compared, because GR and FB give identical output[4]. By contrast, the +LOC dataset, where there are 5 attributes including X-DIM and Y-DIM, and the values of the locative attributes were randomly determined in all domains, there is much greater scope for variation.

All four algorithms also included TYPE by default. Adding TYPE, despite its lack of contrastive value, was the norm in the corpus descriptions (93.5%). While the IA always adds TYPE, as proposed by Dale and Reiter (1995), we applied the same trick to FB and GR to avoid penalising their performance unnecessarily. In addition, we extended each algorithm in two ways:

Plurality: To cover the plural descriptions in the corpus, we used the algorithm of (van Deemter, 2002), which is an extension to the IA. The algorithm first searches through the KB to find a distinguishing conjunction of properties, failing which, it searches through disjunctions of increasing length until a distinguishing description is found. FB and GR can easily be extended in the same way.
Gradable properties: Locative expressions in the corpus are essentially gradable. For instance, *the table on the left* could be used even if the table was located in the right half of the grid, as long as it was the *leftmost* table. van Deemter (2006) proposed an algorithm to deal with such gradable properties, which can use any of the GRE algorithms (FB, GR, IA). Gradable properties are represented in the form $\langle \text{A} = n \rangle$, for example $\langle \text{X-DIM} = 3 \rangle$ (i.e., the property of being located in the middle column of the grid). This equality is converted into a number of inequalities of the forms $\langle \text{X-DIM} > m \rangle$ and $\langle \text{X-DIM} < m' \rangle$. For example, in a domain with 2 objects, in column 2 and 3, this results in the inequalities $\langle \text{X-DIM} > 2 \rangle$ and $\langle \text{X-DIM} < 4 \rangle$. Inequalities are used by a GRE algorithm in the same way as other properties. A postprocessing phase transforms them into superlative form. For example, if a referent is identified by $\langle \text{TYPE} : \textit{sofa} \rangle \wedge \langle \text{X-DIM} > 2 \rangle$, this yields a combination expressible as "the rightmost sofa", or "the sofa on the right".

4.1 Preference orders for the IA

In assessing the impact of preference orders on the IA, we compare some psycholinguistically-motivated versions to a baseline version which reverses the hypothesised trends. In what follows, we denote a preference order using the first letter of the attributes shown in Table 1.

Psycholinguists have shown that attributes such as COLOUR are included in descriptions of objects even when they are not required (Pechmann, 1989; Eikmeyer and Ahlsèn, 1996). Attributes such as SIZE, which require comparison to other objects, are more likely to be omitted (Belke and Meyer, 2002). Based on this research, we hypothesise a 'best' preference order for the IA (IA-BEST_1) in the −LOC dataset, and a reverse baseline order (IA-BASE_1):

IA-BEST_1: C >> O >> S

[4] This is because objects were distinguishable on the basis of three attributes. When only 1 or 2 attributes suffice to distinguish an object, GR and FB always return identical output. In the case of 3 attributes, GR and FB are identical in the present corpus because the minimal description consists of all the properties that have some discriminatory value.

	−LOC			+LOC					
	IA-BEST$_1$	IA-BASE$_1$	GR/FB	IA-BEST$_2$	IA-BEST$_3$	IA-BEST$_4$	IA-BASE$_2$	FB	GR
Mean	.83	.75	.79	.64	.61	.63	.54	.57	.58
Mode	1	.67	.8	.67	.67	.67	.67	.67	.67
PRP	24.1	7.4	18.7	10	4.6	3.9	1.7	6.6	5.8
$\mathbf{t}_S$	7.002^*	-5.850^*	3.333^*	3.934^*	2.313	3.406	.705	.242	.544
$\mathbf{t}_I$	4.632^*	-1.797	1.169	4.574^*	3.352^*	4.313^*	1.776	1.286	1.900

Table 2: Comparison to the Random Baseline ($^*p < .05$)

IA-BASE$_1$: S >> O >> C

In the more complex +LOC dataset, the inclusion of the numeric-valued X-DIM and Y-DIM increases the number of attributes to 5. Arts (2004) found that locatives in the vertical dimension were much more frequent than those in the horizontal (see also Kelleher and Kruijff (2006)). Two different descriptive patterns dominate her data: Either Y-DIM and COLOUR are strongly preferred and X-DIM is strongly dispreferred, or Y-DIM and X-DIM are both highly preferred. This leaves us with three groups of preference orders, namely CY >> {O,S} >> X, YXC >> {O,S}, and Y,C >> {O,S} >> X. Assuming that ORIENTATION preceeds SIZE (which involves comparisons), three promising orders emerge, with a baseline, IA-BASE$_2$, which is predicted to perform much worse.

IA-BEST$_2$: C >> Y >> O >> S >> X

IA-BEST$_3$: Y >> X >> C >> O >> S

IA-BEST$_4$: Y >> C >> O >> S >> X

IA-BASE$_2$: X >> O >> S >> Y >> C

4.2 Differences between algorithms

Table 2 displays mean and modal (most frequent) scores of each algorithm, as well as the *perfect recall percentage* (PRP: the proportion of Dice scores of 1). Pairwise t-tests comparing each algorithm to RAND are reported using subjects (t_S) and items (t_I) as sources of variance. These figures average over all three Cardinality/Similarity conditions; we return to the differences between these below.

With the exception of IA-BASE, the different versions of IA performed best on both datasets. In the simpler −LOC dataset, IA-BEST$_1$ achieved a modal score of 1 24% of the time. Both the modal score and the PRP of GR/FB were lower. Only IA-BEST$_1$ was significantly better than RAND both by subjects and items. This suggests that while IA-BEST$_1$ reflects overall preferences, and increases the likelihood with which a preferred attribute is included in a description, a consideration of the relative discriminatory power of a property, or the overall brevity of a description, does not reflect human tendencies.

A comparison of IA-BEST$_1$ to FB/GR on this dataset showed that the IA was significantly better, though this only approached significance by items. ($t_S = 2.972$, $p = .006$; $t_I(19) = 2.117$, $p = .08$). Though this ostensibly supports the claim of Dale and Reiter (Dale and Reiter, 1995), it should be discussed in the light of the performance of IA-BASE$_1$, which performed significantly *worse* than RAND by subjects, as shown in Table 2, indicating a very substantial impact of the attribute order.

In the +LOC dataset, there is an overall decline in performance. The much poorer performance of FB and GR on this dataset (neither is better than RAND) is due to their not selecting preferred attributes with the same frequency as the better-performing orders of the IA, since the chances of selecting them are contingent on their discriminatory power.

A comparison of GR to FB revealed that the small difference in their mean scores was not significant ($t_1(24) = .773$, ns; $t_2(19) = 1.455$, ns). Pairwise contrasts involving IA-BEST$_2$ showed that it performed significantly better than both FB ($t_S = 4.235$, $p < .05$; $t_I = -2.539$, ns) and GR ($t_S = 4.092$, $p < .05$; $t_I = 2.091$, ns), though only by subjects. This was also the case for IA-BEST$_4$ against FB ($t_S = 3.845$, $p = .01$; $t_I = 2.248$, ns), though not against GR ($t_S = 3.072$, ns; $t_I = 1.723$, ns). None of the comparisons involving IA-BEST$_3$ reached significance. Once again, the performance of the IA on the more complex dataset displays a strong dependency on the predetermined attribute order.

4.3 Plurals and similarity

The final part of the analysis considers the relative performance of the algorithms on singular and plural data, focusing on the best IA in each dataset,

	−LOC		+LOC	
	IA-BEST$_1$	GR	IA-BEST$_2$	GR
SG	.92	.8	.71	.59
PS	.80	.74	.59	.56
PD	.79	.79	.59	.59
F_S	50.367^*	22.1^*	11.098^*	1.893
F_I	40.025^*	2.171	$13.210**$	.611

Table 3: Effect of Cardinality/Similarity $^*p < .001$

IA-BEST$_1$ and IA-BEST$_2$, and on GR (which did not differ from FB in +LOC). As Table 3 shows, the algorithms' performance declined dramatically on the plural data; the difference between the Singular (SG), Plural Similar (PS) and Plural Dissimilar (PD) domains is confirmed by a one-way ANOVA with Cardinality/Similarity as independent variable, though this is not significant for GR in +LOC.

With PS domains (where the minimally required description is always a conjunction), van Deemter's algorithm will succeed at first pass, without needing to search through combinations, except that a disjunction is required for TYPE values (e.g. 2a, below). People tend to be more redundant, because they partition a set if its elements have different values of TYPE, describing each element separately (2b). In the PD condition, the main problem is that the notion of 'preference' becomes problematic once the search space is populated by combinations of attributes, rather than literals.

(2) (a) $\big(desk \vee fan\big) \wedge red \wedge large \wedge forward$

(b) the large red desk facing forward and the large red fan facing forward

5 The People Domain

Like most work in GRE, the preceding results focus on very simple objects, with attributes such as *colour*. With complex objects, the relevant properties are not always easy to ascertain. Similarly, we expect less agreement between corpus annotators and we expect GRE algorithms to perform worse on complex domains, compared to those where objects are simple and stylised. A separate study on the 'people' sub-corpus described in §3 was conducted, using the same overall setup as the present study. In this section, we briefly discuss our main findings in this sub-corpus. A more detailed comparison of the evaluation results on the furniture domain with parallel results on the people domain is reported elsewhere.

The targets in the people corpus differ from their distractors only in whether they had a beard (HASBEARD), wore glasses (HASGLASSES) and/or were young or old (AGE). But as expected, speakers used other attributes than the ones that are necessary to identify the photographed people. As a result descriptions include, for instance, whether a referent wears a tie, or has a certain hairstyle. To be able to match the descriptions with the domain representation a total of 9 attributes were defined per photograph. The first indication that complexity results in much higher variation comes from results on annotator agreement on this data, with the same annotators discussed in §3.3. Though again suggesting a high degree of replicability, the figures indicate greater difficulty in annotating complex data (A: mean = .84, mode = 1 (41.1%); B: mean = .78; mode = 1 (36.3%)).

Another problem is that complex objects, with several attributes, give rise to several possible orders, making it difficult to determine preference orders for the IA *a priori*, particularly since, unlike attributes such as COLOUR and SIZE, there is little psycholinguistic data on reference with attributes such as HASHAIR. Although in the 'people' domain there exists a particular IA algorithm that performs better than the GR algorithm, only a few of the possible preference orders yield significantly better results than GR. When comparing the mean scores of the best IAs from both domains, the best IA in the furniture domain performed much better than the best IA in the people domain. Their mean scores differ substantially: while IA-BEST$_1$ obtained a mean of .83 on furniture descriptions, the best-performing order on the 'people' corpus had a mean of .69, with a lower recall percentage score of 21.3%.

6 Conclusions

In recent years, GRE has extended considerably beyond what was seen as its remit a decade ago, for example by taking linguistic context into account (Krahmer and Theune, 2002; Siddharthan and Copestake, 2004). We have been conservative by focusing on three classic algorithms discussed in Dale and Reiter (1995), with straightforward extensions to plurals and gradables.

We tested the Incremental Algorithm's match against speaker behaviour compared to other models using a a balanced, semantically and pragmatically transparent corpus. It turns out that performance depends on the preference order of the attributes that are used by the IA. Preliminary indications from a study on a more complex sub-corpus

support this view. This evaluation took a *speaker-oriented* perspective. A *reader-oriented* perspective might yield different results. This is our main target for future follow-ups of this work.

One lesson to be drawn from this study is of a practical nature. Suppose a GRE algorithm were required for an NLG system, to be deployed in a novel domain. If the IA is the prime candidate, which preference order should be chosen? Psycholinguistic principles can be good predictors, but an application may involve attributes whose degree of preference is unknown. Investigating how the subjects/authors of interest behave requires time and resources, in the absence of which, an algorithm like GR (suitably adapted to make sure that the TYPE attribute is represented) may be a better bet.

Finding correct preference orders is comparable to a situation wherein a doctor has a choice of two medicines with which to fight the flu. One of these (nicknamed GR) produces reasonable results against all variants of the flu; the success of the other (called IA) depends crucially on a balancing of ingredients that differs from case to case. Finding the right balance is an art rather than a science. – This, we feel, is the situation in GRE today.

7 Acknowledgements

Thanks to Richard Power, Ehud Reiter, Ross Turner, Imtiaz Khan and Emiel Krahmer for helpful comments. This work forms part of the TUNA project (`www.csd.abdn.ac.uk/research/tuna/`), supported by EPSRC grant GR/S13330/01.

References

A. Arts. 2004. *Overspecification in Instructive Texts*. Ph.D. thesis, University of Tilburg.

E. Belke and A. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination. *European Journal of Cognitive Psychology*, 14(2):237–266.

B. Bohnet and R. Dale. 2005. Viewing referring expression generation as search. In *Proc. IJCAI-05*.

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

Robert Dale. 1989. Cooking up referring expressions. In *Proc. ACL-89*.

H. J. Eikmeyer and E. Ahlsèn. 1996. The cognitive process of referring to an object. In *Proc. 16th Scandinavian Conference on Linguistics*.

C. Gardent. 2002. Generating minimal definite descriptions. In *Proc. ACL-02*.

S. Gupta and A. J. Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proc. 1st Workshop on Using Corpora in NLG*.

H. Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. ACL-97*.

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

P. W. Jordan. 2000. Influences on attribute selection in redescriptions: A corpus study. In *Proc. CogSci-00*.

J. D. Kelleher and G-J Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proc. ACL-COLING-06*.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing*. Stanford: CSLI.

T. Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

A. Siddharthan and A. Copestake. 2004. Generating referring expressions in open domains. In *Proc. ACL-04*.

M. Stone. 2000. On identifying sets. In *Proc. INLG-00*.

K. van Deemter, I. van der Sluis, and A. Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proc. INLG-06*.

K. van Deemter. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

K. van Deemter. 2006. Generating referring expressions that contain gradable properties. *Computational Linguistics*. to appear.

I. van der Sluis, A. Gatt, and K. van Deemter. 2006. Manual for the TUNA corpus: Referring expressions in two domains. Technical Report AUCS/TR0705, University of Aberdeen.

J. Viethen and R. Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proc. INLG-06*.