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 focusses 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 Generation of Referring Expressions (GRE) is a core task in most Natural Language Generation systems. The current state of the art in this area is dominated by versions of the Incremental Algorithm (IA) of Dale and Reiter (1995). Focussing 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 focussed 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, abstracting away from other communicative intentions. They 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 which sought to extend the expressiveness and coverage of GRE (Horacek, 1997; Kelleher and Kruijff, 2006). Its original motivations have made it 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 (i.e. not realisation). This is complicated by the semantically intensive nature of GRE. 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. A reliable corpus-based GRE evaluation should ideally be performed against a *semantically and pragmatically transparent* corpus. Semantic transparency means that all the relevant knowledge available to the human authors of the corpus is known. Similarly, pragmatic transparency ensures that the authors’ communicative intentions are known. Ideally, the corpus should be balanced in both respects so that, for example, different kinds

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

of referents occur an equal number of times.

This paper describes the construction of a corpus that meets these requirements, and an evaluation that addressed, as its principal question, 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 Related work

We are aware of three studies that concern the evaluation of GRE algorithms. Two of these (Jordan and Walker, 2005; Gupta and Stent, 2005) compared the IA to some alternative models, using 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. Thus, evaluating the IA against this data may not have done justice to a content determination strategy designed solely to achieve this aim. Furthermore, 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, and compared the IA and GR algorithms against the corpus. However, there are crucial questions that remain unanswered in their study. 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. Moreover, 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.

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 corpus domains

3 A semantically transparent corpus of references

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 descriptons of real photographs of people; the other half contains descriptions of artificially constructed pictures of household items. In this paper, we focus exclusively on the latter sub-corpus. This 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 (or an algorithm) with many alternatives for producing a description. A comparison of the results reported below with parallel results on the people domain can be found in van der Sluis et al. (2007).

3.1 Materials, design and procedure

The household 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 and six distractor objects, placed in a 3 (row) × 5 (column) grid. Pictures of the objects represented combinations of values of the four attributes shown in the top panel of Table 1. In addition, the horizontal and vertical position of the objects is also explicitly represented in domains, 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[2]. 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 told that they had to identify objects for a language-understanding program which would interpret their descriptions and remove the referents from the domain. They were asked to type distinguishing descriptions as though they were interacting remotely with another person. 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. During a debriefing phase, participants who completed the experiment were asked to rate their agreement to the statement *The system performed well on this task.* Of the 5 response categories, ranging from *strongly agree* to *strongly disagree*, 34 individuals selected *agree* or *strongly agree*, while none selected *strongly disagree*.

The corpus is semantically balanced, in that 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):

1. Singular (SG): 7 domains contained a single target referent
2. 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 COLOUR sufficed for a distinguishing description.
3. 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, describing the domain entities, their attribute-value information and location (row and column numbers) in the grid (see Figure 1(a)). 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`. Because of the well-defined nature of the domains, this was only necessary in 39 descriptions (3.2%).

The `DESCRIPTION` tag in Figure 1(b), permits the automatic compilation of a logical form from a human-authored description. 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, as shown in (1).

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

3.3 Annotator reliability

The reliability of the annotation scheme was evaluated in a study involving two independent annotators, 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 annotations of A and B against those by the present authors, using a version of the Dice coefficient of similarity. Let D_1 and D_2 be two descriptions, and $att(D)$ be the attributes in any de-

[2] 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

scription D. The coefficient, which ranges between 0 (no agreement) and 1 (perfect agreement) is calculated as in (2).

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.

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

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.

Both annotators showed a high mean agreement with the present authors in their annotations, as indicated by their mean and modal (most frequent) scores (Annotator A: mean $=$ 0.93, mode $=$ 1 (74.4%); Annotator 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. This measure was adopted 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.

4 Evaluating the algorithms

The three algorithms mentioned in the Introduction 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:

1. **Full Brevity (FB)**: Finds the smallest distinguishing combination of properties.
2. **Greedy (GR)**: Adds properties to a description, always selecting the property with the greatest discriminatory power.
3. **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 corpus was first divided into those descriptions which did not contain locative expressions ($-$LOC dataset) and those which did ($+$LOC dataset). The evaluation was carried out separately for the two datasets. Algorithms were compared to a random baseline (RAND) which picked a property randomly, and added it to the description if it removed distractors and was true of the referents. In the $-$LOC dataset, only GR and IA were compared. This is because objects in this dataset were distinguishable on the basis of three attributes. When MD is of length 1 or 2, GR and FB return identical output. Though this need not be the case when it is of length 3, it is an artefact of the corpus design in the present case, since in these domains, MD consists of all the available attributes (hence every algorithm returns MD). The situation is very different in the $+$LOC dataset, where there are 5 attributes, including X-DIM and Y-DIM, and the minimal description is unpredictable, given that the values of the locative attributes were randomly determined in all domains.

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 had to extend the algorithms in two ways:

1. 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.

2. Gradable properties: Inspection of locative expressions in the corpus revealed that these were essentially gradable. An NP like ‘the table on the left’, for example, was 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) has proposed an algorithm to deal with such gradable properties. The algorithm can use any of the GRE algorithms (FB, GR, IA), as follows. Consider a property of the form $\langle \text{A} = n \rangle$, where n is a real number, for example $\langle \text{X-DIM} = \mathit{3} \rangle$ (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 $\langle \text{X-DIM} > \mathit{m} \rangle$ and $\langle \text{X-DIM} < \mathit{m'} \rangle$. For example, in a domain with 2 objects, in column 2 and 3, this results in the inequalities $\langle \text{X-DIM} > \mathit{2} \rangle$ and $\langle \text{X-DIM} < \mathit{4} \rangle$. The GRE algorithm uses these inequalities in the same way as other properties. In a postprocessing phase, they are transformed into a superlative form. For example, if a referent is identified by $\langle \text{TYPE} : \mathit{sofa} \rangle \wedge \langle \text{X-DIM} > \mathit{2} \rangle$, this yields a combination expressible as “the rightmost sofa”, or “the sofa on the right”.

4.1 Preference orders for the IA

Someone who chooses a preference order to suit a new NLG application does not have to throw a dice: she can consult the psycholinguistic literature. Therefore, when 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 §4.4, we will then consider the extent to which these orders match individual subjects, compared to other possible orders. We denote a preference order using the first letter of the attributes discussed in §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, because they are cognitively more costly (Belke and Meyer, 2002). Based on this research, we hypothesise a ‘best’ preference order for the algorithm (IA-BEST$_1$) in the $-$LOC dataset, and a baseline order (IA-BASE) which reverses it:

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

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 locative expressions in the vertical dimension were much more frequent than those in the horizontal (see also Gapp (1995); 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 (because SIZE 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

	$-$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 ($^*p < .05$)

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[3], indicating a very substantial impact of the attribute order.

In the +LOC dataset, there is an overall decline in the algorithms' performance. The main reason for the much poorer performance of FB and GR on this dataset (neither is better than RAND) is that these algorithms do not select 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.

	$-$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: Mean scores and effect of Plurality/Similarity conditions $^*p < .001$

A comparison of GR to FB in this dataset 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$_3$ showed a significant difference. Once again, the performance of the IA on the more complex dataset displays a strong dependency on the predetermined attribute order; this is supported by the fact that only IA-BEST$_2$ was significantly better than RAND across the board.

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-performing IA in each dataset, and on GR (which was not significantly different 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 MD is always a conjunction), van Deemter's algorithm will succeed at first

[3] This is indicated by the negative value of t_S and t_I.

pass, without needing to search through combinations, except that a disjunction is required for TYPE values (e.g. 3a, below). People tend to be more redundant, because they partition a set if its elements have different values of TYPE, describing each element separately (3b). 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.

(3) (a) $\big(\mathit{desk} \vee \mathit{fan}\big) \wedge \mathit{red} \wedge \mathit{large} \wedge \mathit{forward}$

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

4.4 Differences between subjects

Individual differences are a difficult issue for NLG (Reiter and Sripada, 2002), partly because it is unclear whether NLG systems should use some average of different authors' utterances, or seek to mirror some homogeneous subset of authors. Here, we quantify to what extent subjects differed in terms of what algorithm matches each of them best. We write "s selects algorithm A" as short for "A has the best average match to s's descriptions". Our main question is: *Do different subjects "select" different algorithms?* Restricting attention to the most salient points, we consider the preference orders used earlier, and some other ones (using the same method to compute overall match).

The first step is to find out which algorithms are "selected" by each subject (two or more are selected if they have the same average match with a subjects's descriptions). Consider those 18 subjects who never used location. These selected C>>O>>S most often (16 times), with C>>S>>O in second place (6 times); they selected RAND once. When C>>S>>O or RAND were selected, the difference in Dice scores with C>>O>>S was minimal, at most .02. The difference with the worst-matching version of IA, however, was often substantial, at an average of .084. A typical case is one subject whose match with C>>O>>S was optimal at .87, and with C>>S>>O .86; her worst IA was IA-BASE$_1$ at .76, while GR and FB both scored .83, better than IA-BASE$_1$.

Now consider the 15 subjects who used location consistently. Among these, variation was even greater. As many as 14 different preference orders (possibly *ex equo*) were selected at least once. Among the remaining algorithms, RAND and IA-BASE$_2$ were not selected by any subject, nor were GR and FB. While the version of IA selected by a subject invariably matched its selecting subject quite well (between .64 and .82, with an average of .69), the IA version that matched a subject *worst* was typically dismal (between .31 and .58, with an average of .47), worse than GR (between .46 and .67, with an average of .60).

We conclude that, even in the −LOC data, there are non-negligible differences between subjects. In the +LOC condition, these differences become very substantial. Not only does the difference between the best and worst-matching algorithms become large, but some subjects select algorithms that differ from what psycholinguistic principles predict. (An example of the latter is the order C>>O>>S>>Y>>X, which is selected by 3 out the 15 subjects who used location.)

5 Conclusions

In recent years, GRE has extended considerably beyond what was seen as its remit even ten years ago, for example by taking linguistic context into account (Krahmer and Theune, 2002; Siddharthan and Copestake, 2004). We have been conservative by focussing on three classic algorithms discussed in Dale and Reiter (1995) which are still at the heart of most extensions. Only where extensions of these algorithms were either particularly straightforward (as in the case of simple purals) or necessitated by our experimental setting (as in the case of numeric-valued attributes) did we consider generalisations of these algorithms.

We set out to ask *"Does the Incremental Algorithm* IA *match speakers' behaviour better than other algorithms?* To answer this question, we constructed and annotated a balanced corpus that is semantically and pragmatically transparent, and tested rigourously to what extent each algorithm 'matched' this gold standard. It turns out that the answer depends on the preference order of the attributes that are used by the IA. Our evaluation took a *speaker-oriented* perspective. A *reader-oriented* perspective might yield different results; indeed, 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. Though 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.

Suppose a doctor has a choice of two medicine cocktails with which to fight the flu. One of the cocktails (nicknamed GR) produces reasonable results against all variants of the flu; the success of the other cocktail (called IA) depends crucially on a delicate balancing of ingredients: every epidemic, and every patient, requires a different balance, and finding the right balance is an art rather than a science. – This, we feel, is the situation in GRE today.

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: A comparative study of german and swedish. In *Proc. 16th Scandinavian Conference on Linguistics*.

K.P. Gapp. 1995. Angle, distance, shape, and their relationship to projective relations. In *Proc. CogSci-95*.

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.

E. Reiter and S. Sripada. 2002. Should corpora texts be gold standards for nlg? In *Proc. INLG-02*.

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 (Special Session on Data Sharing and Evaluation)*.

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, University of Aberdeen.

I. van der Sluis, A. Gatt, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. Submitted.

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