Evaluating Algorithms for the Generation of Referring Expressions: Going Beyond Toy Domains

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

Abstract

We describe a corpus-based evaluation methodology, applied to a number of classic algorithms in the generation of referring expressions. Following up on earlier work involving very simple domains, this paper deals with the issues associated with domains that contain 'real-life' objects of some complexity. Results indicate that state of the art algorithms perform very differently when applied to a complex domain. Moreover, if a version of the Incremental Algorithm is used then it becomes of huge importance to select a good preference order, because some preference orders are prone to generating very unnatural output. Finding good preference orders, however, can be difficult, as we show. These results should contribute to a growing debate on the evaluation of NLG systems, arguing in favour of carefully constructed *balanced* and *semantically transparent* corpora.

Keywords

generation of referring expressions, corpus-based evaluation of algorithms, Natural Language Generation

1 Introduction

This paper evaluates some classic algorithms for the Generation of Referring Expressions (GRE), which focus on the question of Content Determination. We ask how well these algorithms model the semantic content of expressions produced by people. It replicates the methodology used in [8], which carried out an evaluation using relatively simple domains of objects with well-defined properties. In addition to presenting new evaluation results on a novel, more complex domain, this paper poses a number of questions regarding the adequacy of existing GRE algorithms when they are deployed in scenarios involving complex objects. In contrast to 'toy' domains, such objects afford human authors with a much larger variety of referential possibilities, with potentially more inter-author variation. This has some consequences for existing GRE algorithms that rely on predefined general or domain-specific 'preferences' for content determination, whereby some properties of objects are prioritised over others. While psycholinguistic research has indeed shown that such preferences exist, the results have tended to rely on precisely the kinds of simple objects that characterise most proposals in the GRE literature. Our aims in this paper are to (a) examine the feasibility of constructing a semantically-annotated corpus for GRE evaluation in complex scenarios; (b) evaluate the performance of current content determination heuristics for GRE on such scenarios; (c) compare this performance to our earlier results on more limited domains.

GRE is a semantically intensive sub-task of microplanning in NLG. GRE algorithms take as input a Knowledge Base (KB), which lists domain entities and their properties (attribute-value pairs), together with a set of intended referents, R. The output is a distinguishing description of R, that is, a logical form which distinguishes this set from its distractors. Most work in the area focuses on very simple objects, with attributes such as *colour* and *shape*. With complex real-world objects, the relevant properties are not always easy to ascertain. For instance, in describing a *person*, attributes such as *shape* become problematic, whereas *hair-colour*, *beard-colour*, *has-glasses* and *height* are not only more relevant, but also more numerous. Some properties (e.g. a person's freckles) may only be used occasionally, or not at all. Because of more variation, GRE algorithms might be expected to perform worse on complex domains, compared to those where objects are simple and stylised. For an evaluation which compares output to the human gold standard represented by a corpus, another problem is the potential for lack of agreement between corpus annotators. This is especially non-trivial since *semantic and pragmatic transparency* are prerequisites for corpora in GRE as we have argued in [23]. 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 of referents occur an equal number of times.

This paper describes the construction of a corpus of this kind, involving a moderately complex domain whose inhabitants are (black & white photographs of) *people*. The resulting corpus is then used to compare some classic GRE algorithms with human descriptions. Wherever appropriate, we shall highlight the ways in which our experiences and findings differed from the ones involving a simpler domain [8].

2 Related work

The current state of the art in GRE is dominated the Incremental Algorithm (IA) of Dale and Reiter [5], which has served as a starting point for later models which sought to extend the expressiveness and coverage of GRE [10, 14, 15, 17, 22]. The IA was proposed as a better match to human referential behaviour relative to some predecessors, notably Dale's [4] Full Brevity (FB) and Greedy (GR) heuristics, which emphasise *brevity* as the main determinant of adequacy. In contrast, the IA performs hillclimbing along a predetermined list of domain attributes. This *preference order* reflects general or domain-specific preferences, which is the main reason for the IA's predicted superiority. However, the preference order strongly impacts the IA's performance, since in a domain with n attributes, there are in principle $n!$ different incremental algorithms.

Few empirical evaluations have been conducted in this area, and those that were done were limited to descriptions of objects that can be identified with only a few clearly distinguishable attributes like *colour* or *type*. [13] and [9] compared the IA to some alternative models, using the COCONUT dialogue corpus, where pieces of furniture are described with four attributes at most. [24] used a small corpus of descriptions of drawers, using *colour* and *location* attributes only. Apart from using simple domains, these studies meet the transparency requirements mentioned above to a very limited degree. Though COCONUT dialogues were elicited against a well-defined domain, [12] 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.

2.1 Computing Similarity

One question that the studies mentioned above raise relates to how human-authored and automatically generated descriptions should be compared. A measure of recall (as used in the Jordan/Walker and Viethen/Dale studies) indicates coverage, 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. To obtain a more fine-grained measure, we use the Dice coefficient of similarity shown in (1). Let D_1 and D_2 be two descriptions, and let $att(D)$ be the attributes in any description D. The coefficient takes into account the number of attributes that an algorithm omits in relation to the human gold standard, and those it includes, making it more optimally informative. Because descriptions could contain more than one instance of an attribute (e.g. 'the young man with the glasses and the old man who also wears glasses'), the sets of attributes for this comparison were represented as multisets.

type	hasBeard	hasGlasses	Age
man	1	0	old
man	1	1	young
man	0	1	old
man	0	0	young

Table 1: *Attributes and example targets as defined in the corpus domains*

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

3 A transparent corpus of references

We constructed and annotated a balanced corpus that pairs each description in the corpus with a logical form that is cast in terms of the domain with respect to which the description was produced. Our corpus contains ca. 1800 descriptions, collected through a controlled experiment run over the web. Participants in the experiment were asked to identify one or two objects from a set of distractors shown on their computer screen, by typing distinguishing descriptions as though they were interacting remotely with another person. One within-subjects variable was the use of different domains: (1) artificially constructed pictures of household items and (2) real photographs of people, yielding two sub-corpora. In this paper, we discuss how the latter sub-corpus is gathered, annotated and used to evaluate various GRE algorithms. Throughout the paper, we compare with our findings on the furniture corpus [8].

3.1 Materials and design

The people sub-corpus consists of 810 descriptions from 45 native or fluent speakers of English. Participants described photographs of men in 18 trials, each corresponding to a domain where there were one or two clearly marked target referents and six distractors (also men), placed in a 3 (row) × 5 (column) grid. The use of these pictures was based on previous experimental work using the same set [19].

In addition to their location (on which more below), all targets could be distinguished via the three attributes shown in Table 1. Thus, the targets differed from their distractors in whether they had a beard (hasBeard), wore glasses (hasGlasses) and/or were young or old (Age). The corpus is semantically balanced, in that for each possible combination of the attributes, there was an equal number of domains in which an identifying description of the target(s) required the use of those attributes (modulo other possibilities). We refer to this as the *minimal description* (MD) of the target set. However, results of earlier studies with the same set of photographed persons indicated that speakers use other attributes to identify the photographed people as well (e.g, whether the person wears a tie, a suit or has a certain hairstyle or colour). These too were included in the corpus annotation, for a total of 9 attributes per photograph.

By contrast, objects in the furniture sub-corpus were invariably described using at most four attributes.

The present study focusses on the subset of the corpus descriptions which do *not* contain locative expressions ($N = 342$ from 19 authors)[1]. For comparison, we use the subset of the household/furniture sub-corpus which also does not contain locatives ($N = 444$ descriptions from 27 authors). Comparing the furniture and people descriptions, the variation amongst the people descriptions is expected to be higher and the annotation of the people descriptions is expected to be more difficult.

The experiment manipulated another within-subjects variable in addition to the domain, namely Cardinality/Similarity (3 levels):

1. Singular (SG): 6 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 wearing glasses in a domain where HASGLASSES='1' sufficed for a distinguishing description.
3. Plural/Dissimilar (PD): 6 Plural trials, in which 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 suggests that they can be treated adequately by minor variations of the classic GRE algorithms ([7, 11]), as long as the descriptions in question refer distributively [20]. This is something we considered worth testing.

3.2 Corpus annotation

To make the corpus semantically transparent, we designed a XML annotation scheme [18] that pairs each corpus description with a representation of the domain in which the description was produced (see Figure 1(a)). In order to match the descriptions produced by the participants in the study with the domain representations, the entities in the people domain are represented with 9 attribute tags in total. Six of them, HASGLASSES, HASBEARD, HASHAIR, HASSHIRT, HASTIE, HASSUIT have a boolean value. The other four attributes have nominal values: the attribute TYPE has values `person` or `other`, the attribute AGE has value `old` or `young`, HAIRCOLOUR has values `dark`, `light` or `other`, and finally ORIENTATION, which captures the gaze direction of a photographed man, has three possible values `frontward`, `leftward` or `rightward`. If a part of a description could not be resolved against the domain representation, it was enclosed in an OTHER attribute tag with the value `other` for `name`. This was necessary in 62 descriptions (18.2%), a figure which is much larger than that obtained in the simpler furniture domain, in which only 3.3% of descriptions contain OTHER tags.

Figure 1(b) shows the annotation of a plural description in the people domain. 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 *with black facial hair* is tagged as HASBEARD, with the value *1*. Note that HASBEARD encloses the HAIRCOLOUR tag used for *black*. 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 (2).[2]

$$\begin{array}{l}([\mathit{Age{:}\ old}] \wedge [\mathit{type{:}\ person}] \wedge \\ [\mathit{Orientation{:}\ frontward}]) \vee ([\mathit{hasBeard{:}\ 1}] \wedge \\ [\mathit{hairColour{:}\ dark}] \wedge [\mathit{type{:}\ person}])\end{array} \quad (2)$$

3.3 Annotator reliability

The reliability of the corpus 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 [18]. They were given a stratified random sample of 540 target descriptions consisting of 270 descriptions from each domain. For both the furniture and the people domain they were given 2 descriptions 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 the Dice coefficient described above. We believe that 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.

Inter-annotator agreement was high in both sub-corpora, as indicated by the mean and modal (most frequent) scores. In the furniture domain, both A and B achieved similar agreement scores with the present authors (A: mean = .93, mode = 1 (74.4%); B: mean = .92; mode = 1 (73%)). They also evinced substantial agreement among themselves (mean = .89, mode = 1 (71.1%)). In the people domain A's annotations were in slightly better agreement with our annotations than B's (A: mean = .84, mode = 1 (41.1%); B: mean = .78; mode = 1 (36.3%)). The annotators had a somewhat higher agreement among themselves than with the annotations of the present authors in the people domain (mean = .89, mode =1 (70%)).

Overall, these results suggest that the annotation scheme used is replicable to a high degree. As expected however, these results also indicate that annotating complex object descriptions is more difficult than ones elicited in simple domains.

[1] Location was manipulated as a between-subjects factor. Participants were randomly placed in groups which varied in whether they could use location or not.

[2] In the phrases of interest, disjunction or set union is the semantic correlate of the use of *and* in a plural description.

```
<ENTITY type=‘target’>
   <ATTRIBUTE name=‘type’ value=‘person’/>
   <ATTRIBUTE name=‘age’ value=‘old’/>
   <ATTRIBUTE name=‘hasBeard’ value=‘0’/>
   <ATTRIBUTE name=‘hasGlasses’ value=‘0’/>
   <ATTRIBUTE name=‘orientation’
value=‘frontward’/>
...
</ENTITY>
 <ENTITY type=‘target’>
   <ATTRIBUTE name=‘type’ value=‘person’/>
   <ATTRIBUTE name=‘age’ value=‘young’/>
   <ATTRIBUTE name=‘orientation’
value=‘frontward’/>
   <ATTRIBUTE name=‘hasBeard’ value=‘1’ />
   <ATTRIBUTE name=‘hasGlasses’ value=‘0’/>
   <ATTRIBUTE name=‘orientation’
value=‘frontward’/>
...
</ENTITY>
```

(a) Fragment of a domain

```
<DESCRIPTION num=‘plural’>
   <DESCRIPTION num=‘singular’>
    <ATTRIBUTE name=‘Age’ value=‘old’>elderly
                            </ATTRIBUTE>
    <ATTRIBUTE name=‘type’ value=‘person’>man
                                 </ATTRIBUTE>
    <ATTRIBUTE name=‘orientation’ value=
    ‘frontward’>who is facing the front
    </ATTRIBUTE>
   </DESCRIPTION>
 and
   <DESCRIPTION num=‘singular’>
    <ATTRIBUTE name=‘type’ value=‘person’>man
                                 </ATTRIBUTE>
    <ATTRIBUTE name=‘hasBeard’ value=‘1’>with
     <ATTRIBUTE name=‘hairColour’ value=‘dark’>
                                  black</ATTRIBUTE>
                        facial hair</ATTRIBUTE>
</DESCRIPTION>
</DESCRIPTION>
```

(b) Example Description

Fig. 1: *Annotation example: ‘elderly man who is facing the front and man with black facial hair’*

4 Evaluating the algorithms

We evaluated the three algorithms introduced earlier, all of which can be characterised as search problems [3]:

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 performance of these algorithms was tested with respect to 342 descriptions in the ‘people’ corpus. Among other things, they were compared to a baseline (RAND), which randomly added properties true of the referent(s) to the description until it was distinguishing. Because the IA always adds TYPE [5], the same trick was applied to all algorithms, to level the playing field.[3]

In addition, we extended the algorithms to cover the plural descriptions in the people corpus, using the algorithm of [21]. This algorithm first searches for a distinguishing description through literals in the KB, failing which, it searches through disjunctions of increasing length until a distinguishing description is found. This approach was applied to FB and GR as well as the different versions of IA.

We had several general expectations regarding this evaluation. In particular, we expected all algorithms to perform worse with respect to the people descriptions than with respect to the furniture descriptions, simply because the larger number of attributes means that there is more room for error. Before we can delve more deeply into these matters, we need to ask what we mean when we speak about the IA, given that this search method gives rise to different algorithms depending on the way in which attributes are ordered.

4.1 Preference orders for the IA

In simple situations, such as the furniture domain in this corpus, which contained 3 attributes (apart from LOCATION), the number of ways in which attributes can be grouped into a preference order for the IA was limited [8]. Psycholinguistic evidence also facilitates the task. For instance, it is known that attributes such as COLOUR tend to be included in descriptions even when they are not required [16, 6, 1], while relative attributes requiring comparison to other objects (such as SIZE), are cognitively more costly and more likely to be omitted [2]. In a more complex domain, such as the people domain in this corpus, the larger number of attributes increases the possible number of preference orders, and testing them all is unfeasible. Moreover, many of these attributes will not have been studied in the psycholinguistics literature. Let us see how these issues pan out in the (only moderately complex!) people domain.

Although the experimental trials on which the people corpus is based were composed in such a way that the targets could be distinguished with a combination of the attributes hasBeard, hasGlasses and AGE, the descriptions contain many other attributes. With the 9 attributes (excluding TYPE) that were needed to annotate the bulk of descriptions in the people corpus, there are as many as $9! = 362880$ possible preference orders. [4] How might one narrow down 362880 prefer-

[3] In the corpus 91% of descriptions in the people domain and 93.5% in the furniture domain contain a TYPE.

[4] For the algorithm evaluation we exclude the OTHER tag, because it represents a variety of unregistered properties.

	mean (SD)	sum
type	1.39 (.64)	475
hasGlasses	.68 (.78)	231
hasBeard	.66 (.56)	226
hairColour	.61 (.54)	210
hasHair	.46 (.62)	158
orientation	.21 (.48)	73
age	.10 (.36)	34
hasTie	.04 (.18)	12
hasSuit	.01 (.11)	4
hasShirt	.01 (.09)	3

Table 2: *Means, Standard Deviations (*SD*), and Sum frequencies of attribute usage in the people domain.*

ence orders to a manageable number? This is evidently an art rather than a science, but it will be instructive to see how one might reason, and how successful or unsuccessful this type of reasoning can be.

Having built the corpus, the natural approach is perhaps to count the frequencies of occurrence of each of the attributes. Table 2 shows that HASGLASSES (G), HASBEARD (B), HAIRCOLOUR (C), and HASHAIR (H), are relatively likely to be included in a description. Arguably, a person's age is an attribute that needs comparison (e.g. with the ages of the distractors), so one might assume that AGE (A) is less preferred than HASGLASSES and HASBEARD.

In the corpus annotation, HAIRCOLOUR can only appear when the HASHAIR or the HASBEARD tag is included in the description (see Section 3). Accordingly, one can reasonably restrict the number of possible preference orders by the constraint that HAIRCOLOUR, can only be positioned in the preference order when preceded by HASHAIR and HASBEARD. The following 8 IAs are tested: IA-GBHC, IA-GHBC, IA-HBGC, IA-HBCG, IA-HGBC, IA-BHGC and IA-BGHC and the 3 algorithms that perform best are presented in the next section. In addition the IA with the worst of all preference orders, IA-WORST, was tested as a baseline case. The preference order used by this algorithm lists the attributes in reverse-frequency order (e.g. HASSHIRT > HASSUIT > HASTIE > AGE > ORIENTATION > HASHAIR > HASBEARD > HAIRCOLOUR > HASGLASSES).

4.2 Differences between algorithms

As indicators of the performance of algorithms, we use mean and modal (most frequent) scores, as well as the *perfect recall percentage* (PRP: the proportion of Dice scores of 1). Pairwise t-tests are used to compare the average Dice scores of each algorithm to RAND and to GR. These comparisons are reported using subjects (t_S) and items (t_I) as sources of variance.

Table 3 displays scores averaged over all three Cardinality/Similarity conditions; we return to the differences between these below. It shows the results of the three best IAs (IA-GBHC, IA-GHBC and IA-BGHC) from the eight IAs that were tested on the people descriptions. Also shown are the results of the IA with the worst preference order IA-WORST as well as the performance of the FB, the GR on the same descriptions. To enable a comparison with the evaluation of algorithms tested on the furniture descriptions the results for GR and for the version of the IA that performed best in this domain are included in the table as well. The latter algorithm, IA-COS, is a version using a preference order consisting of three attributes (COLOUR > ORIENTATION > SIZE).

Results. All eight IA variations that were evaluated with the people corpus perform significantly better than RAND. This baseline achieved a mean Dice score of .47 (SD= .24; PRP=2.6%; Mode= .33). Of the three best IAs shown in Table 3, the algorithm with the highest mean, IA-GBHC, has a modal score of 1 in 21.3% of the cases. (This is also achieved by IA-GHBC.) A pairwise t-test tells us that the IA-GBHC algorithm performs significantly better than IA-GHBC, though its average dice score is only better by subjects ($t_S = 10.720$, $p = .01$; $t_I = 1.678$, *ns*). These figures suggest that even when only the first four attributes in the preference order are varied, differences in performance are already noteworthy. The IA with the worst preference order performs very badly, and much worse than any other algorithm that was considered. Its mean Dice score is .33 and the best match it receives with the descriptions in the corpus is .75, which happened for only one description (thus, its PRP was 0).

The by-subject and by-item analysis for the GR algorithm presented in Table 4 shows that FB performed slightly worse than GR, but only by subjects ($t_S = -4.147$, $p = .01$)). Interestingly, GR also matches the people descriptions better than some of the IAs that were tested. This is most obviously true for IA-WORST, but also for IA-BHCG and IA-HGCB (two of the eight IA algorithms that were tested, but whose values are not shown in Table 3). For instance, IA-BHCG (mean= .60; SD= .21) was significantly worse by subjects than GR ($t_S = 3.187$, $p = .01$; $t_I = 1.159$, *ns*). On the other hand, the average dice scores of GR are significantly lower than the IA that performed best in our analysis, IA-GBHC ($t_S = -3.332$, $p = .01$; $t_I = -3.236$, $p = .01$). These results indicate a very substantial impact of preference orders, which offers a note of caution: in practice, identifying a preference order is not always trivial, and minor variations in attribute orderings can have a significant impact.

Although in the people domain there exists a particular IA algorithm that performs better than the GR algorithm, our findings suggest strongly that only a few of the 362880 IA algorithms render better results than GR. So even though the relative discriminatory power of a property (as used by GR) or the overall brevity of a description (as used by FB) may not exactly reflect human tendencies, these factors are certainly worth considering when one has difficulties in determining a preference order in complex domains like this one. When confronted with a new and 'complex' domain, in which attribute preferences are unknown, a properly modified GR algorithm is a better choice than an arbitrary IA.

Turning to a comparison of furniture and people domains, focusing on the best IAs, their mean scores seem to differ substantially, with IA-COS obtaining .83 on furniture descriptions, compared to .69 obtained by IA-GBHC on the people corpus. Nevertheless, the PRP scores tell a different story: 24.1% on 444 furniture descriptions against 21.3% on 342 people descriptions

	PEOPLE						FURNITURE	
	IA-GBHC	IA-GHBC	IA-BGHC	IA-WORST	FB	GR	IA-COS	GR
Mean (SD)	.69 (.23)	.66 (.25)	.68 (.22)	.33 (.13)	.60 (.27)	.64 (.24)	.83 (.13)	.79 (.16)
Mode	1.00	1.00	.67	.29	1.00	1.00 & .67	1.00	.8
PRP	21.3	21.3	17.3	0.0	19.6	19.3	24.1	18.7
compared to RAND $\mathbf{t}_S$	12.080	12.747	14.737	−12.967	8.397	9.724	7.002	3.333
compared to RAND $\mathbf{t}_I$	8.794	5.642	7.026	−12.254	3.371	5.227	4.632	1.169
compared to GR $\mathbf{t}_S$	−3.332	−3.332	−4.385	15.034	−4.147	–	2.972	–
compared to GR $\mathbf{t}_I$	−1.310	1.310∗	1.582∗	10.007	−1.678∗	–	2.117∗	–

Table 3: *Scores for the three best* IA*s,* IA-WORST*,* FB *and* GR *in the people domain. Related figures for* IA *and* GR *in the furniture domain are included for comparison. Values of $t-$tests by subjects (t_S) and items (t_I) compare each to the Random Baseline* RAND *and to* GR*(*p = not significant, otherwise $p \leq .01$).*

	SINGULARS		SIMILAR PLURALS		DISSIMILAR PLURALS	
	IA-GBHC	IA-COS	IA-GBHC	IA-COS	IA-GBHC	IA-COS
Mean (SD)	.78 (.19)	.92 (.12)	.77 (.22)	.80 (.11)	.51 (.15)	.79 (.13)
Mode	1.00	1	1.00	.8	1.00	.8
PRP	21.3	60.8	21.3	7	21.3	.8

Table 4: *Scores the algorithms as a function of Cardinality/Similarity.*

seem fairly comparable.

One explanation of the overall worse performance of the algorithms on the people domain, which was hypothesised in Section 1, is that there is greater scope for inter-author variation in more complex domains, and perhaps also greater scope for variation within the descriptions produced by the same author. As an approximate indicator of this, we computed the average number of attributes that descriptions in the two domains had. This was clearly higher in the people domain (3.64) than in the furniture domain (2.02). More important than a measure of central tendency however, is the variance. At 2.24, variance in the number of attributes across descriptions of people was substantial, compared to a mere .66 in furniture. This largely confirms our expectations, as well as offering an explanation for some of the different results obtained in the two sub-corpora.

The final part of our analysis concerns the relative performance of the algorithms on singular and plural descriptions. Table 4 displays scores for the best-performing IAs in the furniture and in the people domain as a function of the Cardinality/Similarity variable. Results in the people domain suggest that the algorithm performs approximately equally well in the 'singular' and 'plural similar' conditions. Pairwise comparisons showed no significant difference between these two conditions. The difference between 'singulars' and 'dissimilar plurals' was substantial ($t_S = -14.784$, $p = .01$; $t_I = -8.250\, p = .01$). The same was true of 'similar' and 'dissimilar' plurals ($t_S = -10.773$, $p = .01$; $t_I = -8.701$, $p = .01$). One reason for the worse performance on the 'dissimilar' condition is that here, algorithms needed to use disjunction. Under the generalisation of the IA by [21], this involves searching through disjoined combinations of increasing length, a process which obscures the notion of preference incorporated in the preference order.

A similar analysis by [8] on the different Cardinality/Similarity conditions in the furniture corpus showed a somewhat different picture. All algorithms tested in that paper performed better on singular descriptions, but the difference between 'similar' and 'dissimilar' plurals was not as dramatic. One of the reasons for this has to do with TYPE. In the people domain, all entities had the same value of this attribute (*man*). This means that authors avoided coordination (semantic disjunction) in the 'plural similar' domains, producing descriptions such as *the men with the beard.* In the furniture domains, referents in 'plural similar' domains had different basic-level values of TYPE, and authors were more likely to use disjunction, with descriptions such as *the red table and the red chair.* This interpretation suggests that the basic problem encountered by all algorithms in both domains was with disjunction (which had to be used in the similar cases for furniture descriptions, because of the different values of TYPE).

5 Conclusions

Our study of the *people* domain has significantly reinforced a number of conclusions that we were only able to formulate tentatively when studying the simpler *furniture* domain. In particular:

- As in the furniture domain, the 'best' IA outperformed all other algorithms, but unlike the furniture domain, the 'worst' IA was significantly worse than FB and GR.
- The best IA in the furniture domain performed much better than the best IA in the people domain, although the PRP scores of these algorithms were similar.
- The total number of preference orders for IA was much larger in the people domain than in the furniture domain, and it proved difficult to find efficient ways of zooming in on preference orders that perform well.
- The complexity of the people domain makes itself felt with particular force in the algorithmic performance on dissimilar plurals.

Reflecting on these results, one might argue that the Incremental Algorithm (IA) is not suitably named. IA is not really an *algorithm* but a strategy that can be used by a variety of algorithms and only becomes concrete when a preference order is selected. We showed that, in complex domains, different IAs can perform very differently, so that it is important to distinguish between them and ask which one suits a particular domain and genre best.

What are the practical implications of these results for designing NLG systems to be deployed in novel scenarios? The results of [8] suggested that selecting a preference order matters considerably, even in simple domains. The present work shows that these differences become huge when descriptions of more complex objects are considered. Moreover, psycholinguistic principles are of limited help in selecting a manageable subset of 'promising' preference orders. On the positive side, our results indicate that information about the frequency of occurrence of each attribute in a corpus *can* help. One might, of course, ask how useful this finding is for someone who has not studied the domain/genre before. Such a person, after all, does not possess the corpus to compute the frequencies of attributes. One might hope, however, that a quicker, less controlled experiment would give frequency information that could be used to similar effect, but this is a question for future research.

We have sometimes described the 'people' domain that was studied in this paper as if it were complex. But even though the objects in the domain are messier and more complex than the ones that have figured in most previous studies, calling this domain complex is arguably an overstatement. For example, the domain contains only a limited number of people, and nothing else than people, and that *relations* between people were not even taken into account. One wonders how reasonable preference orders might be chosen in any truly complex domain, how a controlled experiment could be set up in such a domain, or how a workable annotation scheme could be devised for gaining information about speakers' behaviour in such situations. It seems likely to assume that the problems revealed by our study will be even greater in such domains.

References

[1] A. Arts. *Overspecification in Instructive Texts.* PhD thesis, University of Tilburg, 2004.

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

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

[4] R. Dale. Cooking up referring expressions. In *Proc. ACL-89*, 1989.

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

[6] H. J. Eikmeyer and E. Ahlsèn. The cognitive process of referring to an object: A comparative study of german and swedish. In *Proc. 16th Scandinavian Conference on Linguistics*, 1996.

[7] C. Gardent. Generating minimal definite descriptions. In *Proc. ACL-02*, 2002.

[8] A. Gatt, I. van der Sluis, and K. van Deemter. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proc. ENLG-2007*, 2007.

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

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

[11] H. Horacek. On referring to sets of objects naturally. In *Proc. INLG-04*, 2004.

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

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

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

[15] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing*. Stanford: CSLI, 2002.

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

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

[18] I. van der Sluis, A. Gatt, and K. van Deemter. Manual for the TUNA corpus: Referring expressions in two domains. Technical report, University of Aberdeen, 2006.

[19] I. van der Sluis and E. Krahmer. The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proc. ICSLP'04*, Jeju, Korea, 2004.

[20] M. Stone. On identifying sets. In *Proc. INLG-00*, 2000.

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

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

[23] K. van Deemter, I. van der Sluis, and A. Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proc. INLG-06 (Special Session on Data Sharing and Evaluation)*, 2006.

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