Generation of Referring Expressions: Assessing the Incremental Algorithm

Kees van Deemter [1] & Albert Gatt [2]
& Ielka van der Sluis [3] & Richard Power [4]

1 Generation of referring expressions

A substantial amount of recent work in natural language generation has focussed on the generation of "one-shot" referring expressions whose only aim is to identify a target referent. Dale and Reiter's Incremental Algorithm (IA) is often thought to be the best algorithm for maximising the similarity to referring expressions produced by people. We test this hypothesis by eliciting referring expressions from human subjects and computing the similarity between the expressions elicited and the ones generated by algorithms. It turns out that the success of the IA depends substantially on the "preference order" (PO) employed by the IA, particularly in complex domains. While some POs cause the IA to produce referring expressions that are very similar to expressions produced by human subjects, others cause the IA to perform worse than its main competitors; moreover, it turns out to be difficult to predict the success of a PO on the basis of existing psycholinguistic findings or frequencies in corpora. We also examine the computational complexity of the algorithms in question and argue that there are no compelling reasons for preferring the IA over some of its main competitors on these grounds. We conclude that future research on the generation of referring expressions should explore alternatives to the IA, focussing on algorithms, inspired by the Greedy Algorithm, which do not work with a fixed PO.

16 Years ago in a classic article in this journal, Dale and Reiter (1995) introduced the Incremental Algorithm (IA) for the Generation of Referring Expressions (GRE). Although this is the most referred-to publication on GRE so far, there has been surprisingly little work that directly assesses the validity of its claims. For even though a number of empirical studies have been carried out, few of these address the type of reference discussed by Dale and Reiter, as we shall argue. Likewise, Dale and Reiter's arguments concerning the computational complexity of the IA have gone largely unchallenged. The present paper aims to redress this situation, by comparing the IA to its main competitors

[1]University of Aberdeen
[2]University of Malta and Tilburg University
[3]Trinity College Dublin
[4]Open University, Milton Keynes

against data elicited from human participants in a controlled experiment, and by assessing the complexity of the IA. Our main finding is that, in many situations, other, more flexible generation strategies may be preferable to the IA.

Generation of Referring Expressions. GRE algorithms are computational models of people's ability to refer to objects. Reference is a much studied aspect of language, and a key part of communication, which is why GRE algorithms are a key part of almost any Natural Language Generation program (NLG; Reiter and Dale, 2000). Although GRE has the ultimate aim of producing fully fledged referring expressions (i.e., noun phrases), one of its most important subtasks is to determine the *semantic content* of referring expressions. It is on this part of GRE – henceforth called *Content Determination* – that Dale and Reiter focussed, and the same is true of the present paper. Let us sketch what this Content Determination task amounts to. (For details, see section 2.)

Dale and Reiter's position. Appelt and Kronfeld had studied the generation of referring expressions, focussing on a range of difficult issues arising from the fact that referring expressions are an integral part of a larger communicative act. The resulting algorithms generated expressions which took discourse context into account (e.g. Appelt and Kronfeld, 1987; Kronfeld, 1989) and often contained more information than was necessary to identify the referent, for example in order to make it easier for the hearer to carry out the process of identification, or in order to meet additional communicative goals (Appelt, 1985). Although their work offered important insights, its success as a model of human reference was essentially unproven, and this started increasingly to be seen when computational linguists became more data oriented. For this reason, Dale and Reiter decided to re-focus, concentrating on what they saw as the core of the problem, which is to *identify* a referent, focussing on simple situations, where nothing matters except the properties of the referent and of the other objects in the domain (the *distractors*). Other aspects of the utterance context were temporarily disregarded, not because they were deemed unimportant, but because these researchers thought it wise to focus on simple things first. Even though this re-focussing on a narrower view of reference was not universal – some continued to investigate the effects of linguistic context (e.g. Passonneau, 1995; Jordan, 2002; Stone et al., 2003) and the interaction with other communicative goals (e.g. Jordan, 2002; Stone et al., 2003) – it has exerted a strong influence on the direction of computational GRE.

Let us spell out Dale and Reiter's assumptions in more detail. The language-generating program takes as its input a knowledge base (KB) whose content is mutually known by speaker and hearer. The KB ascribes properties to objects, formulated as a combination of an attribute and a value. If the domain objects are pets, for example, then one attribute might be TYPE, with values such as *dog* and *poodle*. The first aim of GRE is to find properties that identify the intended referent uniquely, using a semantic form like ($\langle$TYPE: *poodle*$\rangle$, $\langle$COLOUR: *brown*$\rangle$). Computing such a semantic form is called Content Deter-

mination. The second aim is to express the semantic form in words, for example *the poodle which is brown*.

Dale and Reiter examined a number of algorithms, all of which were based on interpretations of the Gricean Maxims, which revolve around brevity, relevance and truth (Grice, 1975). Roughly speaking, these algorithms sought to minimise the redundancy in a description. In the most extreme case (e.g. Dale, 1989), the Gricean maxims had been interpreted as dictating the choice of the smallest set of properties of a referent that will uniquely identify it. Dale and Reiter contrasted this with a more relaxed interpretation of the maxims, which gave rise to what has come to be known as *the Incremental Algorithm* (IA). They argued that, in the situations envisaged, the IA produces referring expressions that resemble human-generated referring expressions better than its competitors. Following Belz and Gatt 2008, we shall call this the criterion of *human-likeness*. Dale and Reiter also argued that the IA is computationally tractable, and that this constitutes another argument in favour of it. The two arguments can be seen as related if one assumes that neither people nor machines can solve computationally intractable problems in real time.

Aims of the paper. Dale and Reiter's empirical position was based on a reading of the psycholinguistic literature (particularly as summarised in Levelt, 1989). Yet – consistent with existing practice in natural language generation at the time – their paper did not include an empirical test. Later work (Passonneau, 1995; Jordan and Walker, 2000) has tested some of their ideas, as we shall see, but this work has tended to concentrate on different (in fact, more complex) referential situations than the ones on which Dale and Reiter focussed. We aim to put Dale and Reiter's original ideas to the test. Like many other studies, our investigation will focus on Content Determination (unlike (Stone et al., 2003; Krahmer and Theune, 2002)), disregarding words and syntactic constructions. Like Dale and Reiter, we shall focus on the task of identifying the referent, disregarding the well-documented fact that referring expressions can serve other communicative purposes (e.g. Jordan, 2002; Stone et al., 2003), and we shall focus on "one-shot" referring expressions, as produced in a null context where words outside the referring expression do not play a role.

In one respect we have been less conservative. For although Dale and Reiter focussed exclusively on singular descriptions, many references to sets can be generated by variations of the classic GRE algorithms (e.g. van Deemter, 2002; Gardent, 2002; Horacek, 2004; Gatt and van Deemter, 2007). This is something we considered worth testing as part of our general plan. In order not to contaminate the discussion of Dale and Reiter's claims, we only discuss plural references of the kind that can be generated by a relatively simple extension of the IA; moreover, we report results on singulars and plurals separately. For generality, we have studied referring expressions in two different domain types, one of which involves references to furniture (the furniture sub-corpus) and the other to photographs of faces of people (the people sub-corpus).

The plan for the experiment which is the main focus of this paper was first outlined in van Deemter et al. (2006). The first tentative results were reported in Gatt et al. (2007) concerning the furniture sub-corpus, and in van der Sluis et al. (2007) concerning the people sub-corpus. The TUNA corpus and evaluation method have influenced the field considerably, in particular after they were chosen as the basis on which GRE algorithms were evaluated in the First NLG Shared Task and Evaluation Challenge (STEC) on Attribute Selection for Referring Expressions Generation (ASGRE), in the Spring and Summer of 2007[5]. 22 Different algorithms were submitted to this STEC for comparative evaluation, coming from 13 different research teams (Belz and Gatt, 2007). TUNA also featured in a subset of the tasks organised for the second STEC in this area[6], where the number of submitted systems was even greater (Gatt et al., 2008a). As of October 2008, the annotated corpus is available from the Evaluations and Language resources Distribution Agency (ELDA)[7].

The present paper offers a definitive statement of the aims and set-up of the TUNA experiment, the structure of the annotated corpus, and our analyses of the corpus (based on a larger number of subjects, in a greater number of conditions, and validated against a second corpus), superseding our earlier papers on the topic. Consistent with our aim of testing Dale and Reiter's original hypotheses, this empirical analysis is combined with a discussion of the computational complexity of the Incremental Algorithm and its main competitors, and concludes with a discussion of the advantages and disadvantages of incrementality.

2 Some classic GRE algorithms

In order to test Dale and Reiter's claims, we focus our investigation on the algorithms that were most prominently discussed in their paper. We introduce the algorithms briefly before discussing them in more detail.

- Full Brevity (FB). Conceptually this is the simplest approach, discussed in a number of early contributions (e.g. Appelt, 1985), and formalised by Dale (1989). FB minimises the number of properties in the description. Technically, FB is not one algorithm but a class, since minimality can be achieved by many different algorithms. FB is computationally intractable, because in the worst case, the time it takes to find a minimal description grows exponentially with the number of properties that the program can choose between (Reiter, 1990).

- Greedy Algorithm (GR). This faster algorithm was introduced as an approximation to FB by Dale (1989). Rather than searching exhaustively, it selects properties one by one, always choosing the property that is true of the intended referent and excludes the greatest number of distractors. GR does not always produce a minimal description, because a property that

[5] See `http://www.csd.abdn.ac.uk/research/evaluation/` for more information.
[6] See `http://www.itri.brighton.ac.uk/research/reg08/`
[7] See also `http://www.csd.abdn.ac.uk/research/tuna/`.

removes the maximum number of distractors at the time of its inclusion might not remove the maximum number of objects in combination with properties that will later be added.

- Incremental Algorithm (IA). Like GR, this algorithm selects properties one by one, and stops when their combination identifies the referent. Incrementality, in this broad sense which it shares with GR, was earlier hinted at in Appelt (1985) (where it is also pointed out that "Choosing a provably minimal description requires an inordinate amount of effort"). In the IA, the order in which properties are added to the description is not dictated by their discriminatory power, as was the case for GR, but by a fixed PO which tells us, broadly speaking, in which order these properties need to be considered. It is important to realise that this PO is logically distinct from the left-to-right order in which attributes occur within a noun phrase. The idea, supported by psycholinguistic work (Pechmann, 1989, e.g.), is, rather, that some attributes are more prominent in speakers' minds than others, which makes them more likely to be included in descriptions. (The fact that, in human language production, some attributes are selected before others, but realised after them, presents an interesting research problem for psycholinguists (e.g. REF Sedivy et al.) but it goes beyond the scope of the present paper.) Like GR, IA does not allow backtracking: once selected, a property is never withdrawn. As others before us have observed (e.g. Jordan and Walker (2005); Fabbrizio et al. (2008b)), the outcome of the algorithm depends on the choice of PO.

An example. A specially constructed example will clarify how these algorithms can lead to different outcomes. Imagine a small domain of objects $\{a, b, c, d, e, f, g\}$. Now consider the following five attributes, each of which has two values, denoted as val_1 and val_2 . We choose values which are each other's complement. Since complementary values do not overlap, this means that the choice between different values (of a given attribute) is always obvious once the referent is given. Thus, here and elsewhere in this article, we will focus on the question of *attribute* selection, on which most research in this area has concentrated.[8]

ATT_1 $val_1 = \{c, e, f\}$ (poodle), $val_2 = \{a, b, d, g\}$ (chihuahua)
ATT_2 $val_1 = \{a, b, c, e\}$ (black), $val_2 = \{d, f, g\}$ (white)

[8] A different focus would have been possible, for example because one value (e.g., *mammal*) may be more general than another value (e.g., *dog*) of the same attribute. (In section 4.2.2 we shall encounter this phenomenon.) The IA assumes that the different values of an attribute are always roughly equally preferred, with the choice between them depending on matters such as their discriminatory value (i.e., not on a fixed PO). This assumption can be questioned – If an elephant is pink its colour is more worth mentioning than if it is grey – and this can motivate a different IA, whose PO defines an ordering on properties (i.e., combinations of an attribute and a value) rather than on attributes. The choice of the best value of an attribute has, to the best of our knowledge, never been empirically investigated. See Dale and Reiter (1995) (the `FindBestValue` function) for an algorithm and van Deemter (2002) for a logical analysis focussing on problems that arise when the values of an attribute overlap. See also section 8.

ATT_3 $val_1 = \{a, b, e, f\}$ (bastard), $val_2 = \{c, d, g\}$ (thoroughbred)
ATT_4 $val_1 = \{d, e, f, g\}$ (outdoor), $val_2 = \{a, b, c\}$ (indoor)
ATT_5 $val_1 = \{a, b, c, d, e, f\}$ (with tail), $val_2 = \{g\}$ (without tail)

Suppose the target referent is the dog e. FB will combine $\langle$ATT2: *val1*$\rangle$ (black, which is true of the objects $\{a, b, c, e\}$ only) with $\langle$ATT4: *val1*$\rangle$ (outdoor, which is true of $\{d, e, f, g\}$), because these two properties jointly single out e while all other descriptions that manage the same feat happen to contain three or more properties. The output thus consists of the properties *black* and *outdoor*. Notice that realising this content into a description would not be straightforward, since neither property maps naturally to a head noun (it is usually the TYPE of an entity which has this role). This description might be realised as, for example, *the black animal that lives outdoors*, inserting the category or TYPE of the referent (i.e., animal) artificially.

While GR often produces the same description as FB, this is not the case in the present example. GR will start by selecting the property $\langle$ATT1: *val1*$\rangle$ (poodle, corresponding to the set $\{c, e, f\}$), because it is the property that excludes most distractors. Even though only two distractors are left, namely c and f, no single property manages to remove both of these at the same time. As the next step, GR will select either $\langle$ATT2: *val1*$\rangle$ (black, removing f) or $\langle$ATT3: *val1*$\rangle$ (bastard, removing c), but in each case a third property is required to remove the last remaining distractor. GR does not generate the smallest possible description, but something like *the black poodle that lives outdoors* instead.

What description is generated by IA? If the attributes are attempted in the order in which they were listed, so ATT_1 is attempted first, followed by ATT_2, then a version of GR is mimicked precisely. But if ATT_2 is attempted first, followed by ATT_4, then the result is the same as that of FB. In other cases, much lengthier descriptions can result, for example when ATT_5 is attempted first (narrowing down the set of possible referents to $\{a, b, c, d, e, f\}$), and then ATT_2 (narrowing down the set of possible referents to $\{a, b, c, e\}$), followed by ATT_1 (resulting in $\{c, e\}$) and finally ATT_4 (resulting in the set $\{e\}$ which contains only the target referent). The resulting description might be realised as *the black poodle with a tail, which sleeps outdoors*.

This artificial example shows how dramatically the outcome of the IA can depend on PO. These differences make it difficult to test the hypothesis that IA's descriptions resemble human-generated descriptions. Which of all the possible IAs, after all, is the hypothesis about? It is possible that in most actually occurring situations, different POs lead to very similar descriptions, or ones that are similar in terms of their human-likeness. We will show that this is not the case, and that there are important differences between the different versions of IA. This outcome will raise the question how "good" POs might be selected.

The TYPE attribute. Types are usually realised as nouns. POODLE, for example, is a type. Because a referring expression normally requires a noun (we tend to say *the brown dog* even when *brown* would identify the referent), Dale and Reiter gave types special treatment. If TYPE is not selected by the algo-

rithm (e.g., because all objects are of the same type), IA adds the property to the description at the end of the search process. Suppose, for example, the referent is a dog. Now if the IA starts by selecting the property BROWN, and if this property happens to identify the referent uniquely, then the property ⟨TYPE: *dog*⟩ is added, generating *the brown dog*. This makes sure that every description contains one property realisable as a noun. This special treatment of nouns – foreshadowed in Appelt (1985) – can be seen as independent of the search strategy. In fact, we believe that the same considerations that make this a good move in combination with IA make it a good move in combination with other algorithms. In our comparison between algorithms, therefore, we have levelled the playing field by making sure that FB and GR apply the same idea. Each of the algorithms considered in this paper will therefore add a suitable value of the TYPE attribute at the end of the algorithm. This did not significantly change the outcome of any of the algorithms since, in the data we use for our evaluation, TYPE was never either necessary or sufficient to identify a target referent, alone or in combination with other properties (see Section 4.2.3).

Plurals. We decided to examine certain kinds of expressions that refer to *sets* as well. We did this by applying a slightly generalised version of IA to these cases, simplifying a proposal in van Deemter (2002). The original IA stopped adding properties to a description when they singled out the set whose only element is the referent. The new IA does the same, stopping when these properties single out a *set* of referents. The same idea extends naturally to FB and GR. This technique allows us, for example, to use the description (⟨ATT1: *val1*⟩,⟨ATT2: *val2*⟩) to refer to the set $\{c, e\}$. Note that this approach does not work for collective references (e.g. *the parallel lines in this picture*; cf. Stone, 2000), which require drastic departures from the algorithms discussed by Dale and Reiter. Another case which this algorithm will not handle is that requiring plural descriptions involving the union of two different sets, such as *the dogs and the cats* (van Deemter, 2002; Horacek, 2004; Gatt and van Deemter, 2007). These cases too will be omitted from the present study.

3 How to test a GRE algorithm?

Empirical work relevant to GRE falls into two main classes. On the one hand, psycholinguistic studies on reference span at least four decades, since the foundational work of Krauss and Weinheimer (1964). On the other hand, partly as a result of increasing interest in comparative evaluation in Natural Language Generation, a number of studies have compared the IA and some other algorithms against corpus data. We give an overview of the main findings from both types of research.

3.1 Psycholinguistic research on reference

As we have seen, the starting point for contemporary GRE research was the notion of Brevity. Brevity was implicit in an early paper by Olson (1970), who argued that reference has a primarily contrastive function, so that the content of an identifying description would be expected to be determined by the distractors from which the referent is distinguished. A substantial body of evidence now shows that brevity is not the only factor motivating speaker's choices of properties in reference. The phenomenon of *overspecification*, whereby speakers select properties which have little or no contrastive value, was observed in early developmental studies (Ford and Olson, 1975; Whitehurst, 1976; Whitehurst and Sonnenschein, 1978); later studies on adult speakers confirmed the tendency. Pechmann (1984) showed that adult speakers begin to articulate a reference before their scanning of a visual domain is complete, and that perceptually salient attributes, such as COLOUR, were apt to be selected before other properties, such as SIZE, which required more cognitive effort because their value can be determined only by comparison to other objects. Later work showed that TYPE and COLOUR were always used in referring expressions in visual domains, even when they had no contrastive value (Schriefers and Pechmann, 1988; Pechmann, 1989), a result not replicated for SIZE. According to these authors, TYPE has privileged status not only because of syntactic constraints, but because speakers process a referent as a **conceptual gestalt**, central to which is the referent's object class; similarly, they argued that COLOUR forms part of the speaker's mental representation of an object. Similar results have been reported by Mangold and Pobel (1988) and Eikmeyer and Ahlsèn (1996). A slightly different interpretation is given by Belke and Meyer (2002), who interpret the finding in terms of an attribute's *codability*, that is, the ease with which that attribute can be included in a mental representation of an object. Thus, relative attributes such as SIZE are argued to have low codability. Evidence for overspecification has also been found in connection with locative (Arts, 2004) and relational descriptions (Engelhardt et al., 2006).

The demonstration that speakers overspecify has obvious relevance to a study comparing the IA to its predecessors, since one of the consequences of a PO is the potential for referential overspecification, with an increased likelihood that attributes which are given high priority (say, COLOUR), will be used even though they are not required for a distinguishing description.

3.2 Direct comparisons of the IA to other models

One of the first studies to systematically compare the IA to other algorithms focussed on the COCONUT corpus, a collection of task-oriented dialogues in which interlocutors had to resolve a joint task (buying furniture on a fixed budget) (Jordan, 2000a).[9] Jordan's study compared the IA to two models. Jordan's

[9] Also worth mentioning is Passonneau (1995), which focussed on reference in discourse in the context of the *pear stories* of Chafe (1980), where a number of algorithms in the Gricean tradition were compared with approaches based on centering.

Intentional Influences (II) model (Jordan, 2000a, 2002) views content determination as heavily influenced by communicative intentions over and above the identification intention. Thus, the intention to signal agreement with a proposal (in this case, to buy some item of furniture), may motivate the speaker to repeat a description produced by their interlocutor, for example. The authors also included a computational implementation of the *Conceptual Pacts* model proposed by Brennan and Clark (1996), in which a speakers' choice of content for a referring expression is in part based on a tacit 'agreement' with the interlocutor to refer to objects in a specific way. The IA was outperformed by both models. However, the comparison leaves open the question as to whether the IA is an adequate model of referent *identification*, since the models to which it was compared explicitly go beyond this.

Still within a dialogue context, Gupta and Stent (2005) carried out an evaluation on the COCONUT and the MAPTASK (Anderson et al., 1991) corpora. They compared the IA and a version of the Greedy algorithm by Siddharthan and Copestake (2004) against a baseline procedure that included the TYPE of an object, and randomly added further properties until a referent was distinguished. Additionally, each algorithm was augmented with dialogue-oriented heuristics, and coupled with procedures for the realisation of modifiers. Thus, the evaluation used an evaluation metric which combined (a) the degree of agreement between an algorithm's attribute selections and a human's, and (b) the extent to which the automatic realisation of attributes was syntactically a good match to the human realisation.

While these studies offer important insights, they do not directly address the questions outlined in the Introduction to our paper. In the MapTask corpus, for example, most referents are named entities with no true distractors, which can explain why the baseline algorithm outperformed both IA and the Greedy algorithm on this data in Gupta and Stent's study. In the COCONUT corpus, these two algorithms outperformed the baseline, but the original IA was outperformed by variants that incorporated dialogue-oriented heuristics. This is exactly as one would predict, since identification is often not the only referential goal of interlocutors, particularly in the COCONUT corpus, where other factors have been shown to be paramount (Jordan, 2000a; Jordan and Walker, 2005). The evaluation metric used by Gupta and Stent incorporated syntactic factors, going beyond the purely semantic task-definition that the IA sought to address. This, of course, is only a limitation from the viewpoint of a study like the present one, which focusses on Content Determination.

A study by Viethen and Dale (2006) offers a more straightforward comparison of the IA and the Greedy Algorithm. Viethen and Dale stuck to the identification criterion as the sole communicative intention. They tested against a small corpus of 118 descriptions, obtained by asking experimental participants to refer to drawers in a filing cabinet, which differed on four dimensions, namely COLOUR, ROW, COLUMN and whether or not a drawer was in a corner. The primary evaluation metric used was *recall*, defined as the proportion of descriptions in the corpus which an algorithm reproduced perfectly. The comparison of IA and GR revealed a recall rate of 79.6% for the latter, compared to a 95.1%

for the IA (with both figures excluding relational descriptions). Moreover, the corpus contained a limited number (29) of overspecified descriptions, of which the IA reproduced all but five. Although these results seem favourable for the IA, they only tell us which descriptions are generated by *one or more* of all (4! =) 24 possible POs of the attributes. This is because Viethen and Dale combined results from all 24 versions of the algorithm, a legitimate move, but one that obscures the extent to which a given version of the IA, with a particular PO, contributes to the overall rate.

Recently, Di Fabbrizio and colleagues reported on a study that is closely related to the questions of the present paper (Fabbrizio et al., 2008b,a, see Bohnet (2008) for a related approach), comparing different versions of the IA.[10] One version uses a PO that reflects the frequency of attributes in the TUNA corpus as a whole, while the other attempts to model different speakers: when modelling a given speaker, the IA uses a PO which reflects the frequency with which *this speaker* (i.e., this subject in the data collection experiment) used the attribute in question. Although the findings of these studies are intriguing – speaker modelling improved the performance of the algorithm – they need to be treated with some caution. This is because the set of speakers represented in the training set from which speaker-dependent preference orders were obtained was different from the set of speakers represented in the test set on which their algorithms were evaluated. This makes it difficult to interpret the conclusion that 'speaker constraints can be successfully used in standard attribute selection algorithms to improve performance on this task' (Fabbrizio et al., 2008b, p.156). One possible reason for the improvement is that the constraints in question are sufficiently general to apply to *classes* of speakers rather than individual speakers, and hence can be generalised from one sample (individuals represented in the training set) to another (those represented in the test set).

In summary, most existing GRE evaluations do not address the question formulated at the beginning of this paper, because they either placed the IA within the context of a task (such as collaborative dialogue) in which reference is likely to go beyond the primary aim for which IA was designed, or because their evaluation criteria obscure the role of attribute preferences (e.g. by averaging over multiple POs). The work of Dale and Reiter remains central to current work in GRE. To take an example, although the three STECs organised in this area over the past few years have led to novel proposals, with an emphasis on empirical methods Bohnet (2007); Theune et al. (2007); de Lucena and Paraboni (2008); Spanger et al. (2008) and the exploitation of novel frameworks such as genetic algorithms Hervás and Gervás (2009); King (2008), many submissions took the IA or one of its two 'competitors' as a starting point. Gricean Brevity has also been emphasised as a desirable property of algorithms in recent years (Gardent,

[10]The study was carried out in the context of TUNA-REG'08, the second round of Shared Task Evaluation Challenges using this corpus (Gatt and Belz, 2008). For this STEC, two new test sets were generated by partially reproducing the original TUNA methodology. These test sets, which feature in parts of the analysis presented in Section 5 and 6, are briefly described in Section 4. The study by Di Fabbrizio et al. was published after our own studies based on the TUNA corpus described in Section 4 (Gatt et al., 2007; van der Sluis et al., 2007).

2002; Bohnet, 2007). It therefore seems crucial to put the claims made by Dale and Reiter to the test while maintaining their original assumptions.

3.3 Towards an evaluation methodology

The foregoing discussion raises a number of methodological issues that the evaluation experiment reported below seeks to address. First, the IA is a *family* of algorithms, since there are as many versions of it as there are POs. The question then arises as to whether all these possible versions should be considered, with the combinatorial explosion that this brings about. Our approach will be to select only those orders which are 'plausible'. Where possible, we shall attempt to define plausibility in terms of earlier psycholinguistic work. As we have seen in Section 3.1, however, much of this work has focussed on relatively simple, well-defined visual domains with attributes such as COLOUR and SIZE. What of more complex domains in which the variety of attributes increases, and the determination of 'salient' or 'conceptually central' attributes is more difficult? Psycholinguistic research has had less to say about preferences in these contexts. For this reasons, it seemed important to investigate the performance of algorithms in domains that are more complex than the ones that have typically been studied, as well as very simple ones.

Since Dale and Reiter's claims focussed on Content Determination, the aims that we set ourselves suggest that a comparison of GRE algorithms should abstract away from differences in lexical choice and syntactic realisation. Suppose an intended referent has the properties ⟨TYPE: *sofa*⟩ and ⟨COLOUR: *red*⟩, and two human authors produce the descriptions *the settee which is red*, and *the red sofa* respectively. An algorithm which selects both the above properties should be counted as achieving a perfect match to both descriptions. A comparison should also rest on the knowledge that the algorithm and the authors share the same communicative intentions (namely, to identify the referent). Of the studies reviewed above, only Viethen and Dale (2006) and Fabbrizio et al. (2008b) satisfied this requirement. The experiment we employed to collect data aimed to minimise potentially confounding communicative intentions.

In line with Dale and Reiter's starting point, the corpus-based evaluation on which we report here focusses on an assessment of the *humanlikeness* of the descriptions generated by a given GRE algorithm. In other words, we ask how well an algorithm mimics speakers.

3.4 The evaluation metric

An evaluation that compares automatically produced output against human data should take into account partial matches, something that a simple recall measure doesn't do. We therefore adopt the Dice coefficient, a well-accepted distance metric which computes the degree of similarity between two sets in a straightforward way. (Section 8 will briefly discuss alternative metrics.) In keeping will earlier remarks, we shall be assessing the similarity between sets of *attributes* (such as colour), rather than sets of properties (such as green).

The Dice metric is similar to the 'match' metric applied to GRE algorithms by Jordan (2000b) (which was defined as $\frac{X}{N}$, where X is 'the number of attribute inclusions and exclusions that agree with the human data' and N the maximum number of attributes that can be expressed for an entity). Dice is computed by scaling the number of attributes that two descriptions have in common, by the overall size of the two sets:

$$dice(D_H, D_A) = \frac{2 \times |D_H \cap D_A|}{|D_H| + |D_A|} \quad (1)$$

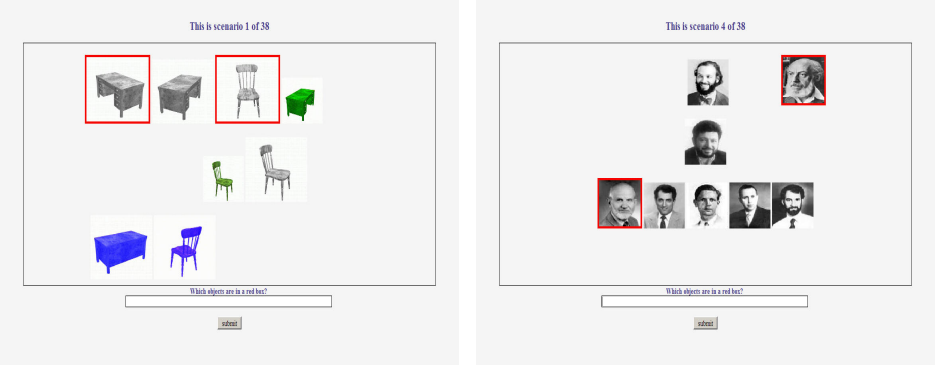
where D_H is (the set of attributes in) the description produced by a human author, and D_A the description generated by an algorithm. Dice yields a value between 0 (no agreement) and 1 (perfect agreement). We will also report the *perfect recall percentage* (PRP), the proportion of times when an algorithm achieves a score of 1, agreeing perfectly with a human author on the semantic content of a description. Finally, a description may contain an attribute several times. For instance, *the green desk and the red chair* contains the attribute COLOUR twice. We treat these as distinct; hence, technically, our Dice coefficient is computed over *multi*-sets of attributes.

4 The TUNA corpus

The experiment that led to the TUNA corpus was carried out over the internet over a period of three months, yielding 2280 singular and plural descriptions by 60 participants. Since this corpus was originally developed, two smaller datasets have been constructed, using the same overall methodology. These were created in 2008 for the comparative evaluation of algorithms in some of the tasks in the second GRE Shared Task Evaluation Campaign, TUNA-REG'08; one of them was also reused in TUNA-REG'09, the third and final GRE STEC involving TUNA held in 2009 (see Gatt and Belz, 2010, for details). The present study compares algorithms against human descriptions in the original TUNA corpus and validates the results against one of the TUNA-REG'08 test datasets. The reason is that the original TUNA data collection incorporated some design features – discussed below – which may have introduced complications in the interpretations of results. Since the REG'08 test sets omitted these features, this comparison gives an additional measure of confidence in our interpretations. We first describe the design, collection and annotation of the original TUNA corpus, followed by a complete overview of the data sources and their use 4.8 below.

4.1 General setup of the experiment

Let us use the term *domain* for a GRE domain in the sense outlined in the previous sections, using *domain type* to refer to the *kinds* of objects in a domain, such as furniture. Different domain types can lead to qualitatively different referring expressions (e.g. Viethen and Dale, 2006; Koolen et al., 2009). We therefore let participants refer to objects in two very different domain types,



(a) Furniture trial (b) People trial

Figure 1: Trials in the TUNA elicitation experiment

yielding two sub-corpora, the *furniture sub-corpus* (Section 4.2.1 and the *people sub-corpus* (Section 4.2.2).

We had to find a setting in which large numbers of reasonably natural human descriptions could be obtained, each referring to an object for the first time (i.e., without being able to rely on previous references to the same referent). These constraints militate against the use of dialogues, which is why we opted for the following, more straightforward approach.[11]

In the experiment, each trial consisted of one or two target referents and six distractor objects, with the targets clearly demarcated by red borders, as illustrated in Figure 1. The participants were asked to identify the objects that were surrounded by the red borders. Participants were told that they would be interacting with a language-understanding program which would interpret their description and remove the referents from the domain. This was intended to introduce a measure of interactiveness into the proceedings. It was emphasised that the aim of these descriptions was to identify the referent. The system was programmed in such a way that one or two objects (depending on whether there were one or two target referents) were automatically removed from the domain after a participant had entered his or her description. To emphasise that this task can be performed with different degrees of success, the system removed the correct referent(s) on 75% of the trials, but the wrong one(s) on a quarter of trials, which were randomly determined.

[11] Koolen et al. (2009) replicated the TUNA experiment with Dutch speakers, manipulating communicative setting: some participants were in a non-dialogue setting while others were in a dialogue setting involving a confederate. In terms of attribute overspecification, no effect of communicative setting was found.

Attribute	Possible values
TYPE	*chair, sofa, desk, fan*
COLOUR	*blue, red, green, grey*
ORIENTATION	*front, back, left, right*
SIZE	*large, small*
X-DIMENSION (column number)	$1, 2, 3, 4, 5$
Y-DIMENSION (row number)	$1, 2, 3$

Table 1: Attributes and values in the *furniture* sub-corpus

Pilots with this scheme suggested that this did not discourage participants from taking the task seriously. Nonetheless, this feature may have introduced confounding factors in our design. For example, participants' referential behaviour may have altered as a result of the system's 'misinterpreting' a description, leading to less risk-taking behaviour (Carletta and Mellish, 1996). This is one of the limitations of this methodology which motivated the use of the TUNA-REG'08 test data as a second, validating test set, since this dataset was constructed without this feature.

The experiment was designed to produce a corpus that would serve as a lasting resource for GRE evaluation; hence, it took into account a number of factors, of which only a few will concern us here. In the following subsections, we describe the materials and design of the experiment, as well as the corpus annotation. A more detailed explanation of the annotation procedure can be found in van der Sluis et al. (2006) and Gatt et al. (2008b).

4.2 Materials

Referential domains (corresponding to experimental trials) consisted of one or two images of target referents and six distractor objects. Objects were displayed in a sparse 3 (row) $\times$ 5 (column) matrix. The positioning of objects was determined randomly at runtime, for each participant and each trial.

4.2.1 The furniture sub-corpus

The furniture sub-corpus describes pictures of furniture and household items obtained from the Object Databank[12], a set of realistic, digitally created images developed by Michael Tarr and colleagues at Brown University. Four types of objects were selected from the Databank, corresponding to four values of the TYPE attribute. For each object, there were four versions corresponding to four different values of ORIENTATION. Pictures were manipulated to create a version of each TYPE $\times$ ORIENTATION combination in four different values of COLOUR

[12] `http://alpha.cog.brown.edu:8200/stimuli/objects/objectdatabank.zip/view`. Compare the COCONUT experiment, where subjects initially saw written descriptions (e.g., "TABLE-HIGH YELLOW $400") instead of pictures.

and two values of size, as shown in Table 1. As shown in the table, there are two additional attributes, X-DIMENSION and Y-DIMENSION, which describe the location of an entity in the 3×5 grid.

4.2.2 The people sub-corpus

The people sub-corpus consists of references elicited in domains consisting of high-contrast, black-and-white photographs of people, following previous experimental work using the same set (van der Sluis and Krahmer, 2004).

Attribute	**Possible values**
TYPE	*person*
ORIENTATION	*front, left, right*
AGE	*young, old*
BEARD	0 (`false`), 1 (`true`), *dark, light, other*
HAIR	0 (`false`), 1 (`true`), *dark, light, other*
HASGLASSES	0 (`false`), 1 (`true`)
HASSHIRT	0, 1
HASTIE	0, 1
HASSUIT	0, 1
X-DIMENSION (column number)	$1, 2, 3, 4, 5$
Y-DIMENSION (row number)	$1, 2, 3$

Table 2: Attributes and values used in the people sub-corpus

This subcorpus is more complex than the furniture one, because a given portrait can be described using a substantial, and perhaps open-ended, number of different attributes (e.g. *the bald man with the friendly smile and the nerdy shirt*). Nevertheless, based on van der Sluis and Krahmer (2004), a number of salient attributes were identified, as shown in Table 2.

The attributes BEARD and HAIR have values which are of *mixed types*. Thus, both can take the boolean values 1 or 0, and either of a set of literal values indicating the colour of a person's hair or beard. In the actual corpus annotation, each of these attributes was a combination of two separate ones, one taking a Boolean value indicating whether a person had the attribute (for example, HASHAIR), and another indicating the colour and (e.g. HAIRCOLOUR). However, the latter was always used in conjunction with the former, in expressions like *dark-haired*. Therefore, it seemed reasonable to combine these into a single attribute in the present study, thereby also reducing the number of attributes overall, and reducing the number of possible POs for the IA in the process.

4.2.3 Construction of domains

The experiment consisted of 38 experimental trials, divided into 20 furniture trials and 18 people trials, each with one or two targets and six distractors in the sparse matrix. Each trial displays a different *domain*. For furniture, the domains were constructed by taking each possible combination of attribute-

value pairs in each domain type[13], and constructing a domain in which that combination was the minimally distinguishing description for the referent(s). For example, in a domain in which the minimal description for the target referent was {⟨COLOUR: *red*⟩, ⟨ORIENTATION: *front*⟩} (*red and facing front*), at least one distractor would be a red chair, and at least one other distractor would be a chair facing front, but only the target referent would have both properties. In the people sub-corpus, the minimal description was calculated based on a combination of three salient attributes found in the van der Sluis and Krahmer (2004) study, namely BEARD, HASGLASSES and the attribute AGE.

The TYPE was never part of the minimally distinguishing description because it was assumed, based on robust psycholinguistic findings, that it would be included anyway. The available attribute-value pairs in a domain type were represented an approximately equal number of times. For example, of the 12 furniture domains where ORIENTATION was part of the minimal description, a target faced *front* or *back* exactly half the time, and *left* or *right* in the rest.

4.3 Design

Table 3 presents an overview of the experimental design, which manipulated one within-subjects, and two between-groups factors. (The abbreviations used in the table will be introduced presently.) The within-subjects factor manipulated the Cardinality and Similarity of objects. In the case of plural domains with two target referents, the two referents may or may not be sufficiently similar to be describable by means of the same minimal conjunction of properties. Where this is not possible, one may have to split the description in two parts (set-theoretically, the union of two sets; logically, the disjunction of two conjunctions of properties) saying, for example, *the red table* and *the blue sofa*. Accordingly, Cardinality/Similarity had three levels:

1. **Singular** (SG): 7 furniture domains, and 6 people domains, contained a single target referent.

2. **Plural/Similar** (PS): 6 furniture domains had two referents with identical values for the attributes with which they could be distinguished from their distractors. For example, two pieces of furniture might both be blue in a domain where the minimally distinguishing description consisted of COLOUR. This was also the case in 6 people domains where, for instance, both targets might be wearing glasses in a domain where ⟨GLASSES: *1*⟩ sufficed for a distinguishing description. In furniture domains, the two referents in this condition had different values of TYPE (e.g. one was a chair, and the other a sofa), while in people domains, they were identical (since all entities were men).

3. **Plural/Dissimilar** (PD): In the remaining 7 plural furniture trials, and the 6 plural people trials, the targets had different values for the minimally distinguishing attributes. Thus, plural descriptions would always

[13]Locative attributes were not used in this calculation, as they were randomly determined.

involve a disjunction (i.e., a set union) if they were to be distinguishing. Since disjunction requires significant extensions to the classic algorithms discussed by Dale and Reiter (1995), we shall omit data in this condition.

	Furniture			**People**		
	SG	PS	PD	SG	PS	PD
+FC-LOC (N = 15)	105	90	105	90	105	105
-FC+LOC (N = 15)	105	90	105	90	105	105
+FC+LOC (N = 15)	105	90	105	90	105	105
-FC-LOC (N = 15)	105	90	105	90	105	105

Table 3: Experimental design and number of descriptions within each cell

The first between-groups factor, ±LOC, consisted in whether participants were encouraged to use locative expressions. Half of the participants were discouraged, though not prevented, from using locative expressions (−LOC condition), while the other half (+LOC) were not. The former were told that the language understanding program they were interacting with had access to the same domain representation, but had different information about the position of objects, so that using locatives would be counter-productive. Participants in +LOC were told that the system had access to the complete domain of objects, including location. Locatives were not included in Dale and Reiter's discussion of GRE algorithms. In recent years, it has become increasingly clear that the location of a referent can play a special role in referring expressions (Arts, 2004), and that location requires special mechanisms for dealing with relations such as *above* (Dale and Haddock, 1991; Kelleher and Kruijff, 2006). Given that the primary focus of this paper is an evaluation of the claims of Dale and Reiter (1995), we do not consider locative descriptions here.

The second between-groups factor sought to determine whether participants would perceive the communicative situation as fault-critical (±FC). The group in the fault-critical (+FC) condition was told that the program would eventually be used in situations where accurate referent identification was crucial, and no opportunity to rectify errors would be available. In this condition, participants could not correct the system's 'mistakes' when it removed the wrong referent(s). Subjects in the −FC condition were not told this. Instead, in the 25% of trials where the program removed the wrong referents, they were asked to click on the correct pictures to rectify the 'error' made by the program. Once again, a full discussion of the effects of the FC factor would take us far beyond the scope of the present paper. We shall therefore collapse descriptions from both ±FC conditions in our analysis based on the original TUNA Corpus. Although some preliminary analysis of the corpus data suggested that there was no difference between the two conditions in the likelihood of overspecified descriptions, some previous work (von Stutterheim et al., 1993; Maes et al., 2004) suggests that manipulation of communicative context may affect referential behaviour. These complications further motivate our validation against the TUNA-REG'08 test data, which did not manipulate a ±FC factor.

```
<TRIAL ID='s2t3' CONDITION='-LOC' CARDINALITY='SG' SIMILARITY='SIMILAR' DOMAIN='furniture'>
    <DOMAIN>
       <ENTITY type='target'>
          <ATTRIBUTE name='type' value='sofa' />
          <ATTRIBUTE name='colour' value='red' />
          <ATTRIBUTE name='orientation'  value='right' />
          <ATTRIBUTE name='size' value='large' />
          <ATTRIBUTE name='x-dimension' value='1' />
          <ATTRIBUTE name='y-dimension'
       </ENTITY>
       <ENTITY type='distractor'>
          <ATTRIBUTE name='type' value='sofa' />
          <ATTRIBUTE name='colour' value='red' />
          <ATTRIBUTE name='orientation' value='left' />
          ...
       </ENTITY>
       ...
    </DOMAIN>

    <DESCRIPTION>
       the
       <ATTRIBUTE ID='a1' name='type' value='sofa'>
          sofa
       </ATTRIBUTE>
       <ATTRIBUTE ID='a2' name='orientation' value='right'>
          facing right
       </ATTRIBUTE>
    </DESCRIPTION>

    <ATTRIBUTE-SET>
       <ATTRIBUTE ID='a1' name='type' value='sofa'/>
       <ATTRIBUTE ID='a2' name='orientation' value='right'/>
    </ATTRIBUTE-SET>
</TRIAL>
```

Figure 2: Example of a corpus instance: "the sofa facing right"

4.4 Participants and procedure

The experiment was run over the internet. Participants were asked for a self-rating of their fluency in English (*native speaker, non-native but fluent, not fluent*). Participants who rated themselves as *not fluent* were not included in the corpus. Participants were then randomly assigned to a condition, and read the corresponding instructions. The instructions emphasised that the purpose of their descriptions is to identify referents. They were asked to complete the experiment (i.e. all 38 furniture and people trials) in one sitting. Trials were presented in randomised order. Each trial consisted of a presentation of a domain, as shown in Figure 1, where participants were prompted for a description of the target referent(s). This was followed by a feedback phase, in which the system removed the target referent. 60 participants completed the experiment, 15 in each group depicted in Table 3.

	Mean (SD)	Mode
Authors vs. A	.886 (.17)	1 (53.2%)
Authors vs. B	.891 (.15)	1 (51.5%)
A vs B	.934 (.15)	1 (72.2%)

Table 4: Mean and modal Dice scores in the inter-annotator reliability study

4.5 Annotation

An XML annotation scheme was developed for the corpus, which pairs each corpus description with a representation of the domain in which it was produced. In the scheme, which is exemplified in Figure 2, a description is represented in three different ways: (a) the original string typed by a participant (the `STRING-DESCRIPTION` node); (b) the same string with all substrings corresponding to an attribute annotated using `ATTRIBUTE` tags (the `DESCRIPTION` node); (c) a simplified representation consisting only of the set of attributes used by a participant (the `ATTRIBUTE-SET` node).

Evaluation required that the domains that were "seen" by humans and algorithms be compatible whenever possible. This was not possible when human-produced expressions contained attributes that were not specified in the domain (e.g., where a person was described as being *serious*); these were tagged using `name='other'`. In Sections 5 and 6, these attributes were treated as different from any system-generated properties.

4.6 Annotation procedure and inter-annotator agreement

The corpus was annotated by two of the authors based on consensus.[14] The reliability of our annotation scheme was evaluated by comparing a subset of 516 descriptions in the corpus to the annotations made by two independent annotators (hereafter A and B), postgraduate students with an interest in NLG, who used the same annotation manual. The Dice coefficient was used as a similarity metric for comparing the three annotated versions of each description. This allows us to measure the *degree* to which annotators agreed on the semantic content of a particular description (Passonneau, 2006, cf.).

Table 4 displays the pairwise mean and modal Dice scores. In all three pairwise comparisons, there is a high degree of similarity, with the most frequent score being 1. However, agreement was slightly higher between the two annotators A and B than between either of them and the authors. To take the likelihood of chance agreement into account, we used Krippendorf's α (Krippendorf, 1980),

[14] van der Sluis et al. (2006) describes the scheme for manual annotation. The annotated text was processed automatically to produce the representation discussed in Section 4.5.

which has the general form shown in 2:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

where D_o is the observed *dis*agreement between annotators, and D_e is the disagreement expected when the units of interest are being coded purely by chance. We follow Carletta (1996) in assuming that a value greater than 0.8 indicates high reliability (see Artstein and Poesio, 2008, for discussion). In the present case, the disagreement on two descriptions D_1 and D_2 is calculated as $1 - dice(D_1, D_2)$. We followed Passonneau (2006) in adopting the following instantiation of (2):

$$\alpha = 1 - \frac{rm - 1}{m} \frac{\sum_i \sum_{D_1, D_2} n_{D_{1_i}} n_{D_{2_i}} 1 - Dice(D_1, D_2)}{\sum_{D_1} \sum_{D_2} n_{D_1} n_{D_2} 1 - Dice(D_1, D_2)} \tag{3}$$

where r is the number of corpus descriptions, m the number of annotators (i.e., 3), i ranges over the individual corpus descriptions and $n_{D_{j_i}}$ is the number of times the set of attributes D_j has been assigned to description i (out of a maximum of 3). The α value obtained was 0.85. This implies that the three sets of independently annotated descriptions were in high agreement and the annotation scheme used is replicable to a high degree.

4.7 A note on the construction of the TUNA-REG'08 datasets

As stated above, we use Test Sets 1 and 2 from the TUNA-REG'08 STEC to further validate our results. They were constructed via an online elicitation experiment based on the original design, with the following differences:

1. Participants were *not* told that they would be interacting with a natural language understanding system, and they received no 'feedback' of the kind given in the original experiment. Rather, they were asked to identify objects as though they were typing them for another person to read.

2. The between-groups factor manipulating whether or not the communicative situation was fault-critical (±FC) was dropped;

3. Only singular domains were used in the experiment. This is because the STECs did not include plural descriptions.

This experiment was completed by 218 participants, of whom 148 were native speakers of English. The test sets were constructed by randomly sampling from the data gathered from native speakers only: both sets contain 112 different domains, divided equally into furniture and people descriptions and sampled evenly from both ±LOC experimental conditions. In Test Set 1, there is one description per domain, for a total of 112 descriptions (56 per domain type). We use this as development data in the present study (sections 5 and 6). In Test Set 2, there are two different descriptions for each domain, for a total of 224 descriptions (112 in each domain type). This constitutes our independent test dataset, used for validation of the results on the original TUNA corpus.

4.8 Summary of testing and development data

Table 5 shows the number of descriptions in our two test sets, and our development set. Because the furniture and people sub-corpora vary so much in complexity, we focus on each one separately in what follows. In each case, the data consists of (a) singular descriptions (SG); and (b) similar plurals (PS) elicited in the −LOC condition. These are the two classes of descriptions that can be handled by the algorithms without extensions to deal with disjunction. Since participants in the −LOC condition were not prevented from using locative attributes (though they were discouraged), we further exclude from our test data all descriptions which include them.

	Source	Cardinality	Furniture	People	Total
Test	Original TUNA Corpus	SG	156	132	288
		PS	158	138	296
	REG'08 Test Set 2	SG	56	56	111
Development	REG'08 Test Set 1	SG	56	56	112

Table 5: Descriptions in the test and development data in the two sub-corpora.

4.9 Algorithms and comparisons

As observed earlier, testing all possible versions of the IA would not be practical, particularly in a domain with a large number of attributes. In each sub-corpus, we therefore selected a subset of the possible IAs, focussing on those which prioritise preferred attributes. Although such preferences can often be identified from previous psycholinguistic work, this is not always possible, especially in the case of the people descriptions. For this reason, we used our development data to estimate frequencies with which different attributes were used. The resulting frequency ranking for each subcorpus was used to determine a set of POs that were predicted to be 'optimal'.

For each subcorpus, we report mean Dice scores for algorithms based on both the original TUNA Corpus and the TUNA-REG'08 test data. For validation purposes, we report correlations between these means. Significant correlations are taken to suggest that the two datasets are compatible, in spite of the methodological differences in the data collection. Focussing on the main test dataset (that is, the original TUNA data), our statistical analysis then proceeds in two steps. First, we compare the set of optimal IAs to the predicted sub-optimal IAs, as well as a random baseline (hereafter referred to as IA-RAND), which always selected the TYPE of an attribute, and then incrementally added randomly-chosen properties to a description until the target referent was identified (this strategy follows Gupta and Stent, 2005). In this part of the study, we are mainly concerned with the impact of different POs on the IA. We then address the question of how the IA compares to other algorithms by identifying the two versions of the IA which are statistically the best and worst, and comparing them to GR

and FB. In each case, we report the results of a by-items ANOVA with Tukey's post-hoc comparisons.

5 The furniture sub-corpus

Our comparison of algorithms begins with the furniture sub-corpus. We first identify candidates for 'plausible' incremental algorithms. As indicated in Section 3.1, there are strong precedents in the psycholinguistic literature for hypothesising that, of the three attributes in this domain, COLOUR will tend to be strongly preferred, while SIZE is dispreferred (e.g. Pechmann, 1989; Belke and Meyer, 2002). The situation is less clear with ORIENTATION.

Attribute	**Frequency (%)**
TYPE	56 (31.6)
COLOUR	49 (27.7)
ORIENTATION	20 (11.3)
SIZE	20 (11.3)

Table 6: Frequency of attribute usage in the development data for the furniture sub-corpus. (Locative attributes and OTHER are omitted because they were ignored in the present study.)

Frequencies computed from our development dataset, displayed in Table 6, confirm the hypothesised trends, but also evince a tie between SIZE and ORIENTATION. We therefore expect that POs that put COLOUR first will generally perform better, but there is the possibility that the relative order of SIZE and ORIENTATION will show a less dramatic difference. We can test these hypotheses by comparing all the possible IAs, as follows:

1. IA-COS: COLOUR >> ORIENTATION >> SIZE
2. IA-CSO: COLOUR >> SIZE >> ORIENTATION
3. IA-OCS: ORIENTATION >> COLOUR >> SIZE
4. IA-OSC: ORIENTATION >> SIZE >> COLOUR
5. IA-SCO: SIZE >> COLOUR >> ORIENTATION
6. IA-SOC: SIZE >> ORIENTATION >> COLOUR

The top panel of Table 7 displays the mean Dice scores for each version of the IA within the two Cardinality conditions under consideration, as well as the PRP obtained by the algorithms across conditions. Overall means are reported for both test sets.

	Original TUNA Data						TUNA-REG'08 Data	
	Singular		**Plural Similar**		**Overall**		**Singular**	
	Mean (SD)	PRP	Mean (SD)	PRP	Mean (SD)	PRP	Mean (SD)	PRP
IA-COS	0.917 (.12)	60.9	0.797 (.10)	7	0.857 (.12)	33.8	0.916 (.16)	69.1
IA-CSO	0.917 (.12)	60.9	0.791 (.11)	7.6	0.853 (.13)	34.1	0.916 (.16)	69.1
RAND	0.840 (.15)	31.4	0.755 (.13)	3.2	0.797 (.14)	17.2	0.826 (.18)	34.6
IA-OCS	0.829 (.14)	25	0.728 (.13)	1.9	0.778 (.14)	13.4	0.829 (.15)	25.5
IA-SCO	0.815 (.14)	19.2	0.730 (.12)	2.5	0.772 (.14)	10.8	0.823 (.15)	18.2
IA-OSC	0.803 (.16)	22.4	0.728 (.13)	1.9	0.765 (.15)	12.1	0.801 (.17)	25.5
IA-SOC	0.780 (.16)	18.6	0.707 (.13)	2.5	0.743 (.15)	10.5	0.782 (.16)	18.2
FB	0.841 (.17)	39.1	0.736 (.14)	4.4	0.788 (.16)	21.7	0.845 (.17)	37.5
GR	0.829 (.17)	37.2	0.721 (.13)	2.5	0.774 (.16)	19.7	0.845 (.17)	37.5

Table 7: Mean Dice scores and standard deviations for the furniture sub-corpus, with PRP scores per algorithm. Plural Similar scores are reported for the original TUNA data only. c, o and s stand for colour, orientation, and size. prp stands for *perfect recall percentage*.

5.1 Comparison of the two test datasets

The mean Dice scores on the two datasets are strongly correlated, both if we compare the overall score on the TUNA data with the TUNA-REG'08 data ($r_9 = .96; p < .001$) and if we compare only the means obtained on the singular descriptions ($r_9 = .985; p < .001$).[15] The ranking of the algorithms is largely the same for the two datasets, particularly for the top and bottom rankings. The overall rankings of the algorithms[16] on the TUNA data also correlated significantly with the rankings on the TUNA-REG'08 data (Spearman's $\rho_9 = .92; p = .001$) as did the rankings obtained on the singular descriptions ($\rho_9 = .98; p < .001$). This suggests that the two datasets are largely compatible, and that the issues raised in relation to the original experimental design did not lead to significant deviations.

5.2 Comparing different versions of the IA

For the next part of our analysis, we focus exclusively on the original TUNA data. As the table shows, all the algorithms performed worse on the plural descriptions, a point to which we return at the end of this section. However, the relative ordering of the different versions of the IA is stable irrespective of the Cardinality condition. Moreover, the trends seem to go in the predicted direction, with the two top IAs being the ones which place COLOUR first, while prioritising SIZE or ORIENTATION leads to a decline. Note also that the PRP,

[15] We compare means for singular descriptions in addition to the overall means on the TUNA data because the TUNA-REG'08 dataset consisted of singulars only.

[16] Each algorithm was assigned a number indicating its rank, where 1 is the top-ranked (best performing) algorithm.

which reflects the extent to which an algorithm agreed perfectly with an individual on a specific domain, declines sharply for those algorithms which do not put the preferred attribute first, while IA-COS and IA-CSO achieve a perfect match with corpus instances more than 60% of the time on the singular data. Moreover, the random baseline (IA-RAND) outperforms all except these two IAs.

A 7 (ALGORITHM) × 2 (CARDINALITY) univariate ANOVA was conducted to compare all the versions of the IA on the original TUNA data. There were highly significant main effects of both factors (ALGORITHM: $F(6, 2184) = 35.67, p < .001$; CARDINALITY: $F(1, 2184) = 291.95, p < .001$). The interaction approached significance ($F(6, 2184) = 2.02, p = .06$).

IA-COS	A		
IA-CSO	A		
IA-RAND		B	
IA-OCS		B	
IA-SCO		B	C
IA-OSC		B	C
IA-SOC			C

Table 8: Homogeneous subsets among versions of the IA in the furniture sub-corpus. Algorithms that do not share a letter are significantly different at $\alpha = .05$.

Pairwise comparisons using Tukey's Honestly Significant Differences yielded the three homogeneous subsets of algorithms (A,B,C) displayed in Table 8. The table evinces a partition between the two IAs that prioritise COLOUR, and the other algorithms. Their performance stands out, in other words, as significantly better than the five other algorithms. They are also the only algorithms that significantly outperform IA-RAND.

These results indicate that even in a small domain with few dimensions of variation among objects, the humanlikeness of the output of the IA is strongly affected by the PO selected. For the next part of our analysis, we will compare GR and FB against an 'optimal' IA, namely IA-COS, and the 'sub-optimal' IA-SOC.

5.3 Comparing the IA to GR and FB

Table 7 (bottom panel) shows that the Greedy and Full Brevity algorithms fall between the two top-scoring IAs which place COLOUR first, and the others, with IA-RAND outperforming them narrowly in terms of mean scores. We conducted a 4 (ALGORITHM) × 2 (CARDINALITY) univariate ANOVA comparing the best and worst IAs to FB and GR. There was a significant main effect of ALGORITHM ($F(3, 1248) = 38.91, p < .001$) and CARDINALITY ($F(1, 1248) = 1.48, p < .001$). Once again, the interaction was not significant ($F(3, 1248) = 1.48, p > .2$). As in the first test, performance on plurals declined for all algorithms.

The results of a post-hoc Tukey's test are presented in Table 9, which shows a clean partition between the brevity-oriented algorithms on the one hand, and the best and worst IAs on the other. (If IA-CSO was used in the comparison,

IA-COS	A		
FB		B	
GR		B	
IA-SOC			C

Table 9: Homogeneous subsets among the best and worst IAs with FB and GR. Algorithms that do not share a letter are significantly different at $\alpha = .05$.

instead of IA-COS, its position would be identical to that of IA-COS in the present table.) No significant difference was obtained between FB and GR; the reason is probably that although a Greedy strategy will not guarantee that the briefest description is found, in a domain with relatively few attributes (and therefore few possibilities) it is likely to converge with a Full Brevity strategy. This is actually supported by the results on the TUNA-REG'08 data, where the two algorithms obtained identical results.

The results in Table 7 show that the humanlikeness of the IA is dependent on the PO. Moreover, the comparison with the other algorithms (Table 9) suggests that the prediction of Dale and Reiter (1995), that an incremental strategy would improve humanlikeness, is only valid for some IAs.

5.4 Discussion

The results so far suggest that the PO is an important component of any analysis that seeks to test Dale and Reiter (1995)'s claims. Having said this, choosing a "good" PO based on psycholinguistic studies would have been easy in the furniture domain type, since these studies suggest that colour is highly preferred Pechmann (1984).

At a more fine-grained level, matters also depend on the evaluation metric used. With the exception of IA-SOC, the overall means in Table 7 range between 0.75 and 0.9. These differences seem intuitively small, an outcome that could be related to the simplicity of the domain. From an engineering point of view (i.e., one that favours robust, feasible solutions that give reasonable results in the long run), the performance of the random baseline IA-RAND, as well as GR and FB, would seem acceptable. On the other hand, the PRP scores distinguish the top-ranking algorithms more sharply.

One outcome which deserves more comment is the difference between singulars and plurals, with plurals yielding consistently lower Dice scores (Table 7). One possible explanation is that the two referents of a plural description in the furniture sub-corpus always had different types. To give the algorithms a fair chance of matching participants' descriptions, we allowed these algorithms to use unions of types, as in 'the blue desk and chair'. It turns out, however, that noun phrases involving unions of types are rare. Instead of 'the blue desk and chair', participants tended to produce descriptions such as 'the blue desk

and the blue chair', which uses the COLOUR attribute redundantly twice. Since our computation of Dice uses multisets, this lowers the overall score of the algorithms. In our example, the human attribute set would contain two occurrences of COLOUR, whereas an algorithm's would contain only one, thus decreasing Dice overall. (In the case of the people sub-corpus, both referents in a plural description had the same TYPE value, since both were always men.)

A full discussion of plurals is beyond the scope of the present paper, but the observations made here do confirm the general thesis that people deviate substantially from Gricean brevity. A closer analysis of plural references suggests, in fact, a substantial impact of the way the objects are categorised (their TYPE) on the form and content of the referring expression.

6 The people sub-corpus

We now turn to the people sub-corpus, where the number of attributes is greater than in our previous analysis. Hence, the possibilities multiply, both in terms of the number of possible versions of the IA, and in terms of the choices the authors had to describe objects.

Compared with the furniture sub-corpus, the larger number of attributes (nine excluding TYPE) in the people sub-corpus makes testing all possible IAs impractical. Therefore, it is even more crucial to have an a priori estimate of the what POs might constitute 'optimal' and 'sub-optimal' IAs. Here, however, the psycholinguistic literature provides less guidance than before. Most of the work cited in Section 3 focussed on references to objects with the kinds of attributes we find in the furniture domain type. (Not a lot has been published on attributes such as HASGLASSES.) Therefore, our reliance on frequencies based on the development dataset is greater than before. Attribute usage frequencies are displayed in Table 10.

Attribute	**Frequency (%)**
TYPE	55 (29.6)
HASBEARD	36 (19.4)
HASGLASSES	25 (13.4)
HASHAIR	22 (11.8)
AGE	14 (7.5)
HASSHIRT	4 (2.2)
HASSUIT	3 (1.6)
HASTIE	2 (1.1)
ORIENTATION	1 (0.5)
X-DIMENSION	12 (6.5)
Y-DIMENSION	12 (6.5)
OTHER	0

Table 10: Frequency of attribute usage in the development data for the people sub-corpus. (Locative attributes and OTHER are ignored in the present study.)

Aside from TYPE, the table suggests a gap between a set consisting of the three attributes BEARD, HASGLASSES and HAIR, and the others. To construct different versions of the IA, we took all possible permutations of these three attributes, imposing a fixed order on the other six. Additionally, we again used a version of the IA that reversed the hypothesised 'best' orders; this is our predicted sub-optimal version. This resulted in the following versions of the IA, in addition to IA-RAND:

1. IA-GBHOATSS:
   HASGLASSES >> BEARD >> HAIR >> ORIENTATION >> AGE >> HASTIE >> HASSHIRT >> HASSUIT
2. IA-GHBOATSS:
   HASGLASSES >> HAIR >> BEARD >> ... >> HASSUIT
3. IA-BGHOATSS:
   BEARD >> HASGLASSES >> HAIR >> ... >> HASSUIT
4. IA-BHGOATSS:
   BEARD >> HAIR >> HASGLASSES >> ... >> HASSUIT
5. IA-HBGOATSS:
   HAIR >> BEARD >> HASGLASSES >> ... >> HASSUIT
6. IA-HGBOATSS:
   HAIR >> HASGLASSES >> BEARD >> ... >> HASSUIT
7. IA-SSTAOHGB:
   HASSUIT >> HASSHIRT >> HASTIE >> AGE >> ORIENTATION >> HAIR >> HASGLASSES >> BEARD

As before, Table 11 gives descriptive statistics for all the algorithms, with the different versions of the IA in the top panel.

6.1 Comparison of the two test datasets

As before, there were strong positive correlations between the means obtained on the two datasets, both when the overall TUNA means are compared to the TUNA-REG'08 means ($r_{10} = .9; p < .001$) and when the singular subset only is compared ($r_{10} = .9; p = .001$). Rankings of algorithms are identical for the singular subset of the TUNA data and the TUNA-REG'08 data. The rankings on the overall TUNA data display some variation in the middle ranks compared to the TUNA-REG'08 data, but the two datasets give the same top and bottom rankings (that is, the top two and bottom two algorithms are the same). This is confirmed by a strong positive correlation between ranks ($\rho_{10} = .88; p = .001$).

	Original TUNA Data				TUNA-REG'08 Data			
	Singular		**Plural Similar**		**Overall**		**Singular**	
	Mean (SD)	PRP	Mean (SD)	PRP	Mean (SD)	PRP	Mean (SD)	PRP
IA-GBHOATSS	0.844 (.17)	44.7	0.819 (.21)	44.9	0.831 (.19)	44.8	0.811 (.17)	33.9
IA-BGHOATSS	0.822 (.17)	36.4	0.776 (.20)	31.9	0.799 (.19)	34.1	0.797 (.17)	32.1
IA-GHBOATSS	0.776 (.21)	29.5	0.759 (.23)	33.3	0.767 (.22)	31.5	0.77 (.18)	26.8
IA-BHGOATSS	0.728 (.19)	15.9	0.683 (.24)	18.1	0.705 (.22)	17	0.792 (.17)	30.3
IA-HGBOATSS	0.688 (.18)	3.8	0.671 (.20)	9.4	0.679 (.19)	6.7	0.765 (.17)	25
IA-HBGOATSS	0.658 (.20)	4.5	0.622 (.22)	6.5	0.640 (.21)	5.6	0.752 (.17)	23.2
IA-RAND	0.598 (.23)	11.4	0.539 (.22)	10.1	0.568 (.23)	10.7	0.527 (.21)	0
IA-SSTAOHBG	0.344 (.11)	0	0.466 (.19)	13.8	0.407 (.16)	7	0.344 (.08)	0
FB	0.764 (.23)	34.1	0.693 (.28)	34.8	0.728 (.26)	34.4	0.642 (.23)	19.6
GR	0.693 (.20)	8.3	0.634 (.23)	10.1	0.663 (.21)	9.3	0.642 (.23)	19.6

Table 11: Mean Dice scores and standard deviations for the people sub-corpus, with PRP scores per algorithm. Plural Similar scores are reported for the original TUNA data only.

6.2 Comparing different versions of the IA

At a glance, the table suggests some important differences between distributions of the scores and algorithm rankings in this sub-corpus compared to the previous one. First, the overall Dice scores are more broadly distributed, with IA-RAND and IA-SSTAOHBG scoring at or below .57. Second, there does not seem to be such a sharp difference in performance between the SG and PS conditions, with only relatively small decreases in performance. The exception is IA-SSTAOHBG, which performs *worse* in the SG condition compared to PS. Third, although IA-RAND once again ranks above the worst-performing IA, the random procedure peforms worse, in relative terms, than it did on the furniture sub-corpus. Fourth, the worst-performing IA has an extremely low PRP of 7, scoring 0 on this measure in the singular data, meaning that it does not achieve a perfect match with any of the descriptions produced by our subjects.

A 6 (ALGORITHM) $\times$ 2(CARDINALITY) univariate ANOVA again showed a significant main effect of ALGORITHM($F(9, 1876) = 118.47, p < .001$) but no main effect of CARDINALITY ($F(1, 1876) = 2.22, p > .1$). However, the interaction was highly significant ($F(6, 1876) = 6.37, p < .001$). These results confirm the preliminary impressions gleaned from the table, where the decline in performance on PS is not as sharp as it was on the furniture data. The interaction is obtained primarily because IA-SSTAOHBG reverses the trend found for all the other algorithms.

Table 12 displays the homogeneous subsets obtained from the Tukey's pairwise comparisons. The table is most interesting in the difference at the two extremes. At the top of the table, the two best-performing algorithms differ significantly from all other algorithms. At the bottom, two distinct subsets

IA-BGHOATSS	A				
IA-GHBOATSS	A				
IA-BHGOATSS		B			
IA-HGBOATSS		B	C		
IA-HBGOATSS			C		
IA-RAND				D	
IA-SSTAOHBG					E

Table 12: Homogeneous subsets among versions of the IA in the people sub-corpus. Algorithms that do not share a letter are significantly different at $\alpha = .05$.

identify the worst-performing algorithms, one of which is IA-RAND. Interestingly, the latter does not cluster with the worst-performing PO. Looking at the two best versions of the IA, the distinction between the POs appears subtle. This is also evident in the overlap between groups B and C, which mirrors the overlap found in the furniture sub-corpus between algorithms that fall between the extremes (Table 8).

6.3 Comparing the IA to GR and FB

As before, the means for GR and FB in Table 7 suggest that they fall somewhere between the best and worst performing IAs. However, whereas their means were not significantly different in the furniture sub-corpus, for the people sub-corpus a significant difference was found, with FB outperforming GR (with a much higher PRP. In this section, we compare these two algorithms to one of the best IAs (IA-BGHOATSS), and the worst (IA-SSTAOHBG).

A 4 (ALGORITHM) $\times$ 2(CARDINALITY) univariate ANOVA revealed the same trends as with the comparison of the IAs in the previous sub-section, with a main effect of ALGORITHM ($F(3, 1072) = 192.63, p < .001$), no main effect of CARDINALITY ($F(1, 1072) = 1.14, p > .25$) and a significant interaction ($F(3, 1072) = 13.44, p < .001$), the latter once again due to IA-SSTAOHBG.

IA-BGHOATSS	A			
FB		B		
GR			C	
IA-SSTAOHBG				D

Table 13: Homogeneous subsets among the best and worst IAs with FB and GR in the people sub-corpus. Algorithms that do not share a letter are significantly different at $\alpha = .05$.

Homogeneous subsets from the pairwise comparisons are shown in Table 13. Apart from confirming the superiority of the top-ranked IA, these results also

confirm that FB outperformed GR, and both are significantly better than the worst version of the IA.

6.4 Discussion

The answers to our questions are generally more clear-cut on this dataset than on the furniture sub-corpus. There is an obvious dependency of the IA on the PO: IA-SSTAOHBG does not achieve a single perfect match with any description, and is significantly worse than all other algorithms, while once again, the versions of the IA which perform best are those which prioritise 'preferred' (i.e., in the present case, frequent) attributes.

The increased number of choices in this domain type also means that a random incremental procedure is more likely to select a distinguishing combination of attributes which a human author would not select. Another observation concerns the distinction between GR and FB; the two algorithms are significantly different on this dataset, with FB performing better and achieving a PRP which approaches that of some of the higher-ranked IAs. In general, GR is likely to overspecify, including attributes that are not minimally required for a distinguishing reference. The fact that FB achieves a respectable performance means that the validity of the generalisation that people are unlikely to be brief is strongly dependent on the domain type.

As we have argued, much of the psycholinguistic literature has shown preferences for attributes such as COLOUR. On one interpretation (Schriefers and Pechmann, 1988; Pechmann, 1989), this is because of a holistic, 'gestalt' mental representation of objects to which some attributes are central owing to their salience and the fact that they are not relative (unlike SIZE). A related interpretation is that these attributes have high codability (Belke and Meyer, 2002). However, predictions of overspecification based on these theories do not seem to carry over straightforwardly to attributes such as whether a person wears glasses, or whether they are bald. The investigation of these phenomena in different domains is an important area of future research.

The main reason why the two Cardinality/Similarity conditions differ less sharply in the people sub-corpus compared to the furniture data is that in the Plural Similar condition in the people domains, the two target referents were both of the same TYPE (i.e., both men). The kinds of 'partitioned' descriptions that were unavoidable in the furniture corpus (because the two referents in a plural target had different values for the TYPE attribute) would sometimes still arise, as in (4a), where different properties were used to characterise each referent. However, where the description ascribes the same properties to each referent as in the PS condition, the form will be as in (4b).

(4) (a) the man with a beard and the man with glasses

(b) the men with beards and glasses

7 Tractability of algorithms

Dale and Reiter claimed that the IA is superior to its competitors in two respects, namely human-likeness and computational tractability. We have so far focussed on the first of these claims, but it is worth discussing the second one as well.

Firstly, suppose we accept tractability as an important consideration. It is then far from clear that this rules out algorithms such as GR, or perhaps even FB. To see why, let us assess the run-time complexity of each of these algorithms, use the following abbreviations:

1. n_a = number of properties known to be true of the intended referent.
2. n_d = number of distractors.
3. n_l = number of attributes mentioned in the final referring expression.

Under Dale and Reiter's analysis, GR has a complexity of $n_a \times n_d \times n_l$, because it needs to make n_l passes through the problem, at each stage checking at most n_a attributes to determine how many of the n_d distractors they rule out. By contrast, the IA has a complexity of $n_d \times n_l$, because it requires n_l passes, but does not look for the optimal attribute at each stage, since this is fixed in the Preference PO. Although this makes GR more computationally 'expensive' than IA, the standard view regarding the complexity of algorithms is that only the general shape of the function matters and not its fine details. Because both algorithms are polynomial, the standard position suggests that they should probably be tarred with the same brush. In other words, there are no strong computational reasons for preferring IA over GR. A worse complexity class is only reached with FB, whose complexity Dale and Reiter assessed as $n_a^{n_l}$ (i.e., exponential).

It's unusual to assess complexity in terms of variables like n_l (which is not known before the end of the calculation) and n_a (which would require an algorithm of its own to calculate). A similar picture emerges, however, if the algorithms are subjected to a more traditional worst-case analysis (van Deemter, 2002). Such an analysis puts the complexity of IA at $n_x \times n_p$, where n_p equals the total number of properties expressible in the language (i.e., the number of attribute/value combinations), and n_x is the number of objects in the domain. In this analysis, it is easy to see that the *order* in which properties are tested is irrelevant, because in the worst case, all properties are tested. The same conclusion follows, namely that both GR and IA have polynomial complexity.[17]

Additionally, is debatable whether it makes sense to dismiss a GRE algorithm purely because it is computationally 'intractable', and it might be partly for this reason that complexity has been discussed relatively rarely in GRE in recent years, when most research has focussed on empirical tests of algorithms on

[17] Both these calculations are formulated in terms of *properties*, as if they were primitive entities rather than combinations of an attribute and a value. While the original IA was incremental in its consideration of *attributes*, this is not the case for its consideration of *values*: given an attribute, the algorithm looked for the most suitable value of this attribute (i.e., using its function `FindBestValue`). See also footnote 8.

miniature domains, as we have seen. Suppose algorithm x produced better output than algorithm y, but at a much slower pace. Would we really want to prefer y over x under all circumstances? The following arguments militate against such a position:

1. Current GRE algorithms do not pretend to model *procedural* aspects of human reference production at all; at best, they offer a good approximation of the descriptions produced by a human speaker. Thus, the primary question that determines a choice between algorithms is which one mimics human output better, not which one is faster.

2. Computational tractability requires one to make assumptions which, in practice, can be debatable. Suppose. for example, no referring expression can contain more than a hundred properties. This reasonable assumption would instantly remove the variable n_l from the formula for the complexity of FB, causing this algorithm to run in polynomial time. An algorithm whose theoretical complexity is polynomial (but whose constants have high values) can easily take more time in practice than an exponential one (whose constants have low values). Thus, it is often difficult to assess the practical implications of complexity results.

To sum up: the experiments reported in previous sections have led us to question the empirical superiority of the IA. Our discussion of computational complexity now tends in the same direction. Combining the evidence, greedy approaches to GRE may have a lot to offer after all.

8 Conclusion

The Incremental Algorithm is by far the best-known GRE algorithm to date. Krahmer and Theune, for example, wrote that 'the Incremental algorithm has become more or less accepted as the state of the art for generating descriptions' (Krahmer and Theune, 2002, p.223). Horacek wrote that 'the incremental algorithm is generally considered best now, and we adopt it for our algorithm, too' (Horacek, 1997, p.207). Recently, Goudbeek et al. (2009) stated that 'Dale and Reiter's (1995) Incremental Algorithm is often considered the algorithm of choice (...), due to its algorithmic simplicity and empirical groundedness'. Oberlander stated that the IA 'clearly achieves reasonable output much of the time' (Oberlander, 1998, p.506). The IA has also become the basis of much recent work that seeks to widen the coverage of GRE algorithms. Examples include work on relational descriptions (e.g. Areces et al., 2008), salience (such as Piwek (2009), which takes the incremental model 'as a point of departure'), vague descriptions van Deemter (2006), and spatial descriptions (e.g. Kelleher and Kruijff, 2006; Turner et al., 2007). No other GRE algorithm can boast a similar popularity, particularly for identifying a referent in a 'null context' (cf. the Introduction section). Similarly, the FB and GR strategies are still among

the main competitors of the IA. Perhaps the main other contender is the graph-based approach of Krahmer et al. (2003), but its main selling point is arguably that it can encode a wide variety of algorithms, including IA, FB and GR.

We have not been able to confirm the advantages that have been claimed for the IA. From a point of view of run-time complexity, there are no strong reasons for preferring IA over the Greedy Algorithm, and in the corpora that we studied, it would be misleading to say that *the* IA matches human-produced descriptions more closely; for although there always existed a version of the IA that outperformed all other algorithms examined, this is not surprising given the fact that, in simple domain types, any halfway reasonable description can be produced by the IA if a preference order (PO) is hand-picked.[18] The success of the IA depended substantially on PO: a sub-optimal PO produces descriptions that are worse than FB and GR. This was not only true when "unreasonable" PO were used, but also when all available evidence, including corpus frequencies, were taken into account to find a good PO. Furthermore, as we hope to explain in more detail elsewhere, an analysis of differences between participants in our experiment revealed that some human speakers are modelled more accurately by FB and GR than via any incremental generation strategy: in the people data, FB agreed perfectly with a human author about 61% of the time, for example. Combining all the evidence, we conclude that someone who is looking for a GRE algorithm for a previously unstudied application domain might do better choosing GR (augmented with a Dale and Reiter-style treatment of head nouns, as we have argued in section 2), instead of an unproven version of the IA.

Because this paper has used the assumptions outlined in Dale and Reiter (1995), our evaluations have focussed on the extent to which the descriptions produced by an algorithm matched human-produced descriptions. It is intrinsically interesting to construct algorithmic models of human language production, for example because it can help computer programs to, one day, pass the Turing test (Turing, 1950). Moreover, there is some evidence that, by making referring expressions resemble the ones produced by human speakers, the resulting expressions tend to be more easily understood (Campana et al., 2004). For all these reasons, human-likeness is currently the prevalent perspective on evaluation of GRE algorithms, including the recent STEC challenges (Gatt and Belz, 2010). Having said this, we acknowledge that there should also be room for alternative, utility-driven evaluation methods. For, although some psycholinguistic theories emphasise the cooperative nature of reference (e.g. Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996), there is evidence that producers do not necessarily maintain a model of a receiver's communicative needs (e.g. Keysar et al., 2003; Engelhardt et al., 2006; Arnold, 2008).

The shortcomings of the IA became particularly noticeable in connection with the more complex of the two domain types, which involved black-and-white photographs of people's faces (i.e., the people domain type). But our people domains were still comparatively simple. Real people would have been

[18] See section 2 for illustration. The claim in the text holds only for uniquely identifying descriptions that do not contain any logically superfluous properties.

identifiable in terms of their physical features, past actions, and so on. It is unclear whether any of the algorithms discussed here would do well in such situations. Some studies, in fact, suggest that there are domain types in which the IA performs poorly regardless of PO.[19] These results suggest to us that future research in GRE should pay close attention to the complexities posed by the large and complex domains that speakers are faced with in real life.

Research on GRE has moved on considerably since 1995, when Dale and Reiter put forward their hypotheses (see Krahmer and van Deemter (2011) for a survey). Yet, the core GRE problem of producing naturalistic "one-shot" descriptions of a single referent continues to attract considerable interest, as was demonstrated by the recent GRE evaluation campaigns. The investigation on which we have reported in this paper raises the question whether incrementality is basically on the right track, or whether some other approach to reference generation is perhaps superior. It appears to us that a nuanced answer to this question is called for, since the choice depends on the situation in which the algorithm has to operate.

In situations where it is possible to see, based on experimental evidence for example, that certain attributes are preferred over others (e.g., because they are easier to perceive or to express), the IA has considerable appeal, because this algorithm allows us to translate this evidence directly into a PO.[20] In situations where neither intuitions nor experimental evidence are available, a version of the Greedy Algorithm is likely to be a better choice than the IA.

The main strength of the Greedy Algorithm lies in its ability to determine the usefulness of a property dynamically: in some cases, *size* will have great discriminatory power, for example, but if 90% of the domain elements have the same size as the target referent then size will be nearly useless for identifying the target. The Greedy Algorithm is able to take such differences into account, because it selects properties based on how many distractors they remove.

But arguably, discriminatory power is not enough. Suppose only one person in the room has green eyes. This does not necessarily make eye colour very useful for referring to this person, because the colour of someone's eyes is difficult to perceive from a distance. A defender of the Greedy Algorithm might counter that if eye colour is *im*perceptible, this attribute should not be available to the generator: GRE algorithms should only use properties that are common knowledge. The IA does have a subtle advantage in such cases, since its PO allows one to order attributes according to their *degree* of perceptibility, avoiding a sharp, and ultimately arbitrary, distinction between perceptible and

[19] An example is Paraboni et al. (2007), where reference in complex, hierarchically structured domains was studied. The authors found that no single IA was able to generate the types of elaborate descriptions ('picture 5 in section 3', rather than 'picture 5', where the latter was minimally distinguishing) that were most preferred by both authors and readers. Other algorithms, designed to produce carefully over-specified descriptions, proved superior to the IA in such situations.

[20] The same is true if the evidence is different across the different values of an attribute, in which case the IA will operate with a PO defined over properties (i.e., combinations of an attribute and a value) instead of attributes. See also footnote 8.

imperceptible. What the IA cannot do, however, is make eye colour more highly preferred for referents nearby than for referents further afield: it must always be preferred to the same degree. The fact that different POs can be selected for different domain types and text genres does not alter this.

We want to make a plea for algorithms that determine *dynamically* which attributes to select for inclusion into a description, based on features of the situation. Discriminatory power can play a role, but so can the extremity of a property (see below; also van Deemter (2006), section 4.1), intentional influences (Jordan, 2000b,a), and alignment (Goudbeek et al., 2009). (A framework suitable for combining several factors into one "cost" is Krahmer et al. (2003).) To show what we have in mind, let us say a bit more about one factor.

The idea of the extremity of a property can be traced back to some well-designed but little-known experiments (Hermann and Deutsch, 1976). When participants were shown pairs of candles, which differed in height and width, it turned out that when asked to refer to a candle that was both taller and fatter than its distractor, speakers overwhelmingly did this by expressing whichever of the two dimensions made the target "stand" out more: when the target candle had a width of 50mm and a height of 120mm, for example, while the distractor had a width of 25mm and a height of 100mm, speakers referred to the target as "dick" (fat), rather than "lang" (tall), because the relative difference between 50mm and 25mm is greater than between 120mm and 100mm. Since both properties (i.e., width and height) would have ruled out the only available distractor, their discriminatory power *in this situation* is equal, yet there was a tendency for speakers to express the more extreme property. Results of this kind are difficult to replicate using a fixed PO. The study of GRE algorithms that avoid a rigid preference order and select their properties "dynamically" – making use of a combination of discriminatory power, extremity, and other factors – appears to us to be one of the most promising directions of work on GRE at the moment.

Acknowledgment: The bulk of the work reported in this article was carried out under the TUNA project (2003-2007), which was supported by EPSRC grant no. GR/S13330/01.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34:351–366.

Appelt, D. (1985). Planning english referring expressions. *Artificial Intelligence*, 26(1):1–33.

Appelt, D. and Kronfeld, A. (1987). A computational model of referring. In *Proc. of 10th Int. Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 640–647.

Areces, C., Koller, A., and Striegnitz., K. (2008). Referring expressions as formulas of description logic. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG-08)*.

Arnold, J. E. (2008). Reference production: Production internal and addressee-oriented processes. *Language and Cognitive Processes*, 23(4):495–527.

Arts, A. (2004). *Overspecification in Instructive Texts*. PhD thesis, Univ. of Tilburg.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics (survey article). *Computational Linguistics*, 34(4):555–596.

Belke, E. and Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

Belz, A. and Gatt, A. (2007). The attribute selection for gre challenge: Overview and evaluation results. In *Proc. of UCNLG+MT: Language Generation and Machine Translation*.

Belz, A. and Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proc. of 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*.

Bohnet, B. (2007). IS-FBN, IS-FBS, IS-IAC: The adaptation of two classic algorithms for the generation of referring expressions in order to produce expressions like humans do. In *Proc. of the Language Generation and Machine Translation Workshop (UCNLG+MT) at MT Summit XI*.

Bohnet, B. (2008). The fingerprint of human referring expressions and their surface realization with graph transducers. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG-08)*.

Brennan, S. and Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

Campana, E., Tanenhaus, M., Allen, J., and Remington, R. (2004). Evaluating cognitive load in spoken language interfaces using a dual-task paradigm. In *Proc. of 9th Int. Conference on Spoken Language Processing (ICSLP'04)*.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, J. and Mellish, C. (1996). Risk-taking and recovery in task-oriented dialogues. *Journal of Pragmatics*, 26:71–107.

Chafe, W. L., editor (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Ablex, Norwood, NJ.

Clark, H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22:1–39.

Dale, R. (1989). Cooking up referring expressions. In *Proc. of 27th Annual Meeting of the Association for Computational Linguistics, ACL-89*.

Dale, R. and Haddock, N. (1991). Generating referring expressions containing relations. In *Proc. of 5th Conference of the European Chapter of the Association for Computational Linguistics*.

Dale, R. and Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*,

19(8):233–263.

de Lucena, D. and Paraboni, I. (2008). USP-EACH frequency-based greedy attribute selection for referring expressions generation. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG'08)*, pages 219–220.

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

Engelhardt, P. E., Bailey, K., and Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54:554–573.

Fabbrizio, G. D., Stent, A. J., and Bangalore, S. (2008a). Referring expression generation using speaker-based attribute selection and trainable realization (ATT-REG). In *Proc. of 5th Int. Conference on Natural Language Generation (INLG'08)*, pages 211–214.

Fabbrizio, G. D., Stent, A. J., and Bangalore, S. (2008b). Trainable speaker-based referring expression generation. In *Proc. of 12th Conference on Computational Natural Language Learning (CONLL'08)*, pages 151–158.

Ford, W. and Olson, D. (1975). The elaboration of the noun phrase in children's object descriptions. *Journal of Experimental Child Psychology*, 19:371–382.

Gardent, C. (2002). Generating minimal definite descriptions. In *Proc. of 40th Annual Meeting of the Association for Computational Linguistics, ACL-02*.

Gatt, A. and Belz, A. (2008). The tuna challenge 2008: Overview and evaluation results. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG-08)*.

Gatt, A. and Belz, A. (2010). Introducing shared task evaluation to nlg: The TUNA shared task evaluation challenges. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*. Springer.

Gatt, A., Belz, A., and Kow, E. (2008a). The tuna challenge 2008: Overview and evaluation results. In *Proc. of 5th Int. Conference on Natural Language Generation, INLG-08*.

Gatt, A. and van Deemter, K. (2007). Incremental generation of plural descriptions: Similarity and partitioning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP-07*.

Gatt, A., van der Sluis, I., and van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proc. of the 11th European Workshop on Natural Language Generation, ENLG-07*.

Gatt, A., van der Sluis, I., and van Deemter, K. (2008b). Xml format guidelines for the tuna corpus. Technical report, Computing Science, Univ. of Aberdeen, http://www.csd.abdn.ac.uk/∼agatt/home/pubs/tunaFormat.pdf.

Goudbeek, M., Krahmer, E., and Swerts, M. (2009). Alignment of (dis)preferred properties during the production of referring expressions. In *Proc. of Workshop "Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference" (PRE-CogSci 2009)*.

Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics: Speech Acts.*, volume III. Academic Press.

Gupta, S. and Stent, A. J. (2005). Automatic evaluation of referring expression

generation using corpora. In *Proc. of 1st Workshop on Using Corpora in NLG, Birmingham, UK.*

Hermann, T. and Deutsch, W. (1976). *Psychologie der Objektbenennung*. Huber, Bern.

Hervás, R. and Gervás, P. (2009). Evolutionary and case-based approaches to REG. In *Proc. of 12th European Workshop on Natural Language Generation (ENLG-09).*

Horacek, H. (1997). An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of 35th Annual Meeting of the Association for Computational Linguistics, ACL-97.*, pages 206–213, Madrid.

Horacek, H. (2004). On referring to sets of objects naturally. In *Proc. of 3rd Int. Conference on Natural Language Generation, INLG-04.*

Jordan, P. and Walker, M. (2000). Learning attribute selections for non-pronominal expressions. In *Proc. of 38th Annual Meeting of the Association for Computational Linguistics.*

Jordan, P. W. (2000a). Influences on attribute selection in redescriptions: A corpus study. In *Proc. of the Cognitive Science Conference.*

Jordan, P. W. (2000b). *Intentional Influences on Object Redescriptions in Dialogue: Evidnece from an Empirical Study.* PhD thesis, Univ. of Pittsburgh.

Jordan, P. W. (2002). Contextual influences on attribute selection for repeated descriptions. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Reference and Presupposition in Natural Language Generation and Understanding.* CSLI Publications, Stanford, Ca.

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

Kelleher, J. D. and Kruijff, G.-J. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proc. of joint 21st Int. Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL/COLING-06.*

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89:25–41.

King, J. (2008). OSU-GP: Attribute selection using genetic programming. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG-08).*

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2009). Need I say more? On factors causing referential overspecification. In *Proc. of Workshop "Production of Referring Expressions: Bridging Computational and Psycholinguistic Approaches" (*PRE-COGSCI'09*).*

Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation.* Stanford: CSLI.

Krahmer, E. and van Deemter, K. (2011). Computational generation of referring expressions: a survey. *in preparation.*

Krahmer, E., van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Krauss, R. and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114.

Krippendorf, K. (1980). *Content Analysis*. Sage Publications, Newbury Park, Ca.

Kronfeld, A. (1989). Conversationally relevant descriptions. In *Proc. of 27th Annual Meeting of the Association for Computational Linguistics, ACL-89*.

Levelt, W. M. J. (1989). *Speaking: From Intention to Articulation.* MIT Press.

Maes, A., Arts, A., and Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, 37(2):117–144.

Mangold, R. and Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, 7(3–4):181–191.

Oberlander, J. (1998). Do the right thing ... but expect the unexpected. *Computational Linguistics*, 24(3):501–507.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77:257–273.

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Computational Linguistics*, 33(2):229–254.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. of 5th Int. Conference on Language Resources and Evaluation, LREC-2006*.

Passonneau, R. J. (1995). Integrating Gricean and attentional constraints. In *Proc. of 14th Int. Joint Conference on Artificial Intelligence (IJCAI-95)*.

Pechmann, T. (1984). Accentuation and redundancy in children's and adults' referential communication. In Bouma, H. and Bouwhuis, D. G., editors, *Attention and Performance*, volume 10. Lawrence Erlbaum, Hillsdale, NJ.

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

Piwek, P. (2009). Salience and pointing in multimodal reference. In *Proc. of Workshop "Production of Referring Expressions: bridging the gap between computational and empirical approaches to generating reference" (PRE-CogSci'09)*.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proc. 28th Annual Meeting of the Association for Computational Linguistics*.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Schriefers, H. and Pechmann, T. (1988). Incremental production of referential noun phrases by human speakers. In Zock, M. and Sabah, G., editors, *Advances in Natural Language Generation*, volume 1. Pinter, London.

Siddharthan, A. and Copestake, A. (2004). Generating referring expressions in open domains. In *Proc. of 42nd Annual Meeting of the Association for Computational Linguistics, ACL-04*.

Spanger, P., Kurosawa, T., and Tokunaga, T. (2008). TITCH: Attribute se-

lection based on discrimination power and frequency. In *Proc. of 5th Int. Conference on Natural Language Generation (INLG-08).*

Stone, M. (2000). On identifying sets. In *Proc. of 1st Int. Conference on Natural Language Generation, INLG-00.*

Stone, M., Doran, C., Webber, B., Bleam, T., and Palmer, M. (2003). Microplanning with communicative intentions: The SPUD system. *Computational Intelligence*, 19(4):311–381.

Theune, M., Touset, P., Viethen, J., and Krahmer, E. (2007). Cost-based attribute selection for generating referring expressions (GRAPH-FP and GRAPH-SC). In *Proc. of UCNLG+MT: Language Generation and Machine Translation*, pages 95–97.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX(2236):433–460.

Turner, R., Sripada, S., Reiter, E., and Davy, I. (2007). Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In *Proc. of the Conference on Applications and Innovations in Intelligent Systems XV (AI-07)*, Cambridge, UK.

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

van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proc. of 4th Int. Conference on Natural Language Generation (Special Session on Data Sharing and Evaluation), INLG-06.*

van der Sluis, I., Gatt, A., and van Deemter, K. (2006). Manual for the tuna corpus: Referring expressions in two domains. Technical report, Univ. of Aberdeen., http://www.csd.abdn.ac.uk/∼ivdsluis/TunaCorpusManual/.

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

van der Sluis, I. and Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proc. of ICSLP-2004, Octobre 4-8, Jeju, Korea.*

Viethen, J. and Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proc. of 4th Int. Conference on Natural Language Generation, INLG-06.*

von Stutterheim, C., Mangold-Allwinn, R., Barattelli, S., Kohlmann, U., and Kölbing, H.-G. (1993). Reference to objects in text production. *Belgian Journal of Linguistics*, 8:99–125.

Whitehurst, G. and Sonnenschein, S. (1978). The development of communication: Attribute variation leads to contrast failure. *Journal of Experimental Child Psychology*, 25:490–504.

Whitehurst, G. J. (1976). The development of communication: Changes with age and modeling. *Child Development*, 47(473–482).