# Generating Referring Expressions that involve Gradable Properties

Kees van Deemter*
University of Aberdeen

*This paper examines the role of gradable properties in referring expressions, from a perspective of natural language generation. Firstly, we propose a simple semantic analysis of* vague descriptions *(i.e., referring expressions that contain gradable adjectives) that reflects the context-dependent meaning of the adjectives in them. Secondly, we show how this type of analysis can inform algorithms for the generation of vague descriptions from numerical data. Thirdly, we ask when such descriptions should be used. The paper concludes with a discussion of* salience *and* pointing*, which are analysed as if they were gradable adjectives.*

**1. Introduction: Vagueness of Gradable Adjectives**

**Vague descriptions.** Vague or gradable expressions pose problems to models of language, caused by their context-dependence, and by the fact that they are applicable to different degrees. This paper focuses on gradable *adjectives*, also called degree adjec-

* Computing Science Department, King's College, University of Aberdeen, United Kingdom, E-mail: kvdeemter@csd.abdn.ac.uk

Submission received: 7th July 2004; Revised submission received: 19th October 2005; Accepted for publication: 24th November 2005

tives.[1] More specifically, we shall explore how referring expressions containing gradable adjectives can be produced by a Natural Language Generation (NLG) program. Following Pinkal (1979), such expressions will be called *vague descriptions* even though, as we shall see, the vagueness of the adjective does not extend to the description as a whole. It will be useful to generalise over different forms of the adjective, covering the superlative form (e.g., 'largest') and the comparative form ('larger'), as well as the positive or base form ('large') of the adjective. Vague descriptions are worth studying because they use vagueness in a comparatively transparent way, often combining clarity of reference with indeterminacy of meaning; as a result, they allow us to make inroads into the difficult area of research on vagueness. *Generation* offers an interesting perspective because it forces one to ask when it is a good idea to use these descriptions, in addition to asking what they mean.

Gradability is especially widespread in adjectives. A search of the British National Corpus BNC, for example, shows at least seven of the ten most frequent adjectives (last, other, new, good, old, great, high, small, different, large) to be gradable. Children use vague adjectives among their first dozens of words (Peccei 1994) and understand some of their intricacies as early as their 24th month (Ebeling and Gelman 1994). These intricacies include what they call Perceptual context-dependence, as when a set of objects is perceptually available and the adjective is applied to an element or subset of the set (e.g., 'Is this hat big or is it little?', when two hats of different sizes are visible).

**Vagueness in** NLG**.** Some NLG systems produce gradable adjectives. The FOG weather-forecast system, for example, uses numerical input (`Rain(Tuesday) = 45mm`) to generate vague output ('Heavy rain fell on Tuesday', Goldberg 1994). FOG does not appear to have generic rules governing the use of gradable notions: it does not compute the meaning of a vague term based on the context, but uses fixed boundary values instead. A more flexible approach is used by Reiter and Sripada (2002), where users can specify boundary values for attributes like rainfall, specifying for example that rain counts as *moderate* above 7mm p/h, as *heavy* above 20mm p/h, and so on. A third approach was implemented in *Dial Your Disc* (DYD), where the extension of a gradable adjective like 'famous' was computed rather than specified by hand (van Deemter and Odijk 1997). To determine, for example, whether one of Mozart's piano sonatas could be called 'a famous sonata', the system looked up the number $x$ of Compact Disc recordings of this sonata (as listed in an encyclopaedia) and compared it to the average number $y$ of CD recordings of each of Mozart's sonatas. The sonata was called a famous sonata if $x >> y$. Like DYD, the work reported in this paper will abandon the use of fixed boundary values for gradable adjectives, letting these values depend on the context in which the adjective is used.

Sometimes we are *forced* to be vague because our information itself (e.g., based on perception or verbal reports) is inexact. Such cases can be modelled by letting NLG systems take vague information (e.g., `Rain(Wednesday) = heavy`) as their input. We shall focus on the more challenging case where the output of the generator is less precise than the input, as is the case in FOG and DYD. This can be a hazardous affair, since vague expressions tend to be interpreted in different ways by different people (Toogood 1980),

1 We take such adjectives to be ones that have comparative and superlative forms, and which can be premodified by intensifiers such as 'very' (Quirk et al. 1972, section 5.4).

sometimes in stark contrast with the intention of the speaker/writer (Berry et al. 2002). We shall therefore focus – unlike earlier computational accounts – on vague *descriptions*, that is, vague expressions in definite descriptions. Here, the context tends to obliterate the vagueness associated with the adjective. Suppose you enter a vet's surgery in the company of two dogs: a big one on a leash, and a tiny one in your arms. The vet asks 'Who's the patient?', and you answer 'the big dog'. This answer will allow the vet to pick out the patient just as reliably as if you had said 'the one on the leash'; the fact that 'big' is a vague term is irrelevant. You omit the exact size of the dog, just like some of its other properties (e.g., the leash), because they do not improve the description. This shows how *vague* properties can contribute to the *precise* task of identifying a referent.

**Plan of this paper.** We will show how existing algorithms for the Generation of Referring Expressions (GRE) can do justice to gradable properties, whether they originate from the gradable adjectives in a vague description, or from some entirely different source (such as the degree of salience of the referent). Considerable attention will be paid to the many open questions in this area, which will have to be resolved before NLG can be said to contain a proper treatment of vague expressions. We start with two preliminary sections, containing a semantic analysis of vague descriptions (section 2) and a version of the Incremental Algorithm that generates references to sets (section 3). Section 4 describes the core of one particular algorithm for generating vague descriptions. Section 5 discusses pragmatic constraints that let such an algorithm avoid descriptions that are semantically correct but clumsy. Section 6 discusses Linguistic Realisation. Section 7 summarises some empirical results. Section 8 explores non-incremental versions of our algorithm. Section 9 shows how our approach can be extended to include nouns, salience, and pointing. Section 10 sums up our main findings.

## 2. The Meaning of Vague Descriptions

**Linguistic motivation.** We shall be studying vague descriptions of various forms: they may or may not contain a numeral *n* (positioned before or after the adjective); and the gradable adjective (Adj) may at least be in base ('large') or superlative form ('largest'):

1. *The (n) Adj(est) N* (singular/plural)
2. *The Adj(est) (n) N* (singular/plural).

If *Adj* is in the base form, we focus on the word order (1); if *Adj* is superlative, we focus on (2). (Little will hinge on this decision.) We are limiting ourselves to *referential* uses of these expressions, excluding cases like 'This must be *the largest tree in the world*', in which the expression ascribes a property to an already-identified object. Likewise, we exclude intensional ones (e.g., 'Consider *the smallest element of this set*', in a mathematical proof, when the identity of the element may not be known).

Many different analyses are possible of what it means to be large: larger than average, larger than most, larger than some given baseline, and so on. It is doubtful that any one of these analyses makes sense for all definite descriptions. To see this, consider a domain of three mice, sized 5cm, 8cm, 10cm.[2] Here one can speak of

---

2 The reader is asked to focus on any reasonable size measurement, e.g., the maximal horizontal or vertical distance, or some combination of dimensions (Kamp 1975; also section 8.1 of the present paper).

3. *The large mouse* (= the one whose size is 10cm).
4. *The two large mice* (= the two whose sizes are 8 and 10cm).

Clearly, what it takes for the adjective to be applicable has not been cast in stone, but is open to *fiat*: the speaker may decide that 8cm is enough, or she may set the standards higher (cf., Kennedy 1999.) The numeral (whether it is implicit, as in (3), or explicit) can be construed as allowing the reader to draw inferences about the standards employed (Kyburg and Morreau 2000, DeVault and Stone 2004): (3), for example, implies a standard that counts 10cm as large and 8cm as not large. Our own proposal will abstract away from the effects of linguistic context. We shall ask how noun phrases like the ones in (3) and (4) can be generated, without asking how they constrain, and are constrained by, other uses of 'large' and related words. This will allow us to make the following simplification: in a definite description that expresses only properties that are needed for singling out a referent, we take the base form of the adjective to be *semantically* equivalent with the superlative form (and, analogously, the comparative):

*The n large mice = The largest n mice*
*The large mice = The largest mice*
*The large mouse = The largest mouse*.

Viewed in this way, gradable adjectives are an extreme example of the 'efficiency of language' (Barwise and Perry 1983): far from meaning something concrete like 'larger than 8cm' – a concept that would have very limited applicability – or even something more general like 'larger than the average N', a word like 'large' is applicable across a wide range of different situations.

**Caveat: Full NP anaphora.** Having said this, there are *pragmatic* differences between the base form and the superlative (section 5). For example, the equivalence does not take anaphoric uses into account, such as when 'the large mouse' is legitimised by the fact that the mouse has been called 'large' before, as in

5. *I was transfixed by* a large mouse *on the chimney; then suddenly, dozens of mice were teeming on the ground.* The large mouse *was running away.*,

where the mouse on the chimney may be smaller than those on the ground. We focus on Ebeling and Gelman's (1994) *perceptual* context dependence (section 1), pretending that the only contextually relevant factor is the 'comparison set': those elements of the noun denotation that are perceptually available. We disregard *functional* context-dependence, as when 'the small hat' is the one too small to fit on your head.

**Caveat: Evaluative adjectives.** What we wrote has also disregarded elements of the 'global' (i.e., not immediately available) context. For some adjectives, including the ones that Bierwisch called *evaluative* (as opposed to *dimensional*), this is clearly inadequate. He argued that evaluative adjectives (such as 'beautiful' and its antonym 'ugly'; 'smart' and its antonym 'stupid', *etc.*) can be recognised by the way in which they compare with antonyms. For example (after Bierwisch 1989),

6a. 'Hans is taller than Fritz' $\Rightarrow$ 'Fritz is shorter than Hans'.
6b. 'Hans is smarter than Fritz' $\not\Rightarrow$ 'Fritz is more stupid than Hans'.

We could require that the referent of an evaluative description falls into the correct segment of the relevant dimension. (For Fritz to be 'the stupid man', it is not enough for him to be the least intelligent male in the local context; he also has to be a fairly stupid specimen in his own right.) If this is done, it is not evident that dimensional adjectives should be treated differently: If Hans' and Fritz' heights are 210cm and 205cm respectively, then it seems questionable to describe Fritz as 'the short man', even if Hans is the only other man in the local context (but see Sedivy et al. 1999, discussed in section 7.2). Be this as it may, we shall henceforth focus on local context, assuming that additional

requirements on the global context can be made if necessary.

With these qualifications in place, let us say more precisely what we will assume the different types of expressions to mean. For ease of reading, concrete examples (e.g., 'large') will replace abstract labels (e.g., 'Adj'), but the analysis is meant to be general.

> **'The largest *n* mouse/mice'; The *n* large mice.** Imagine a set $C$ of contextually relevant animals. Then these noun phrases (NPs) presuppose that there is a subset $S$ of $C$ that contains $n$ elements, all of which are mice, and such that (1) $C - S \neq \emptyset$ (i.e., not all elements of $C$ are elements of $S$) and (2) every mouse in $C - S$ (i.e., every contextually relevant mouse not in $S$) is smaller than every mouse in $S$. If such a set $S$ exists then the NP denotes $S$. The case where $n = 1$, realised as 'The large(st) mouse', falls out automatically.
>
> **'The large(st) mice'.** This account can be extended to cover cases of the form 'The [Adj]-(est) [$N_{pl}$]' ($pl$ = plural), where the numeral $n$ is suppressed: they will be taken to be ambiguous between all expressions 'The [Adj]-(est) $n$ [$N_{pl}$]', where $n > 1$. Sometimes, this leaves only one possibility. For instance, in a domain where there are five mice, of sizes 4, 4, 4, 5, and 6cm, the only possible value of $n$ is 2, causing the NP to denote the two mice of 5 and 6cm size.

Pragmatic refinements are discussed in section 5. Our analysis is limited to NPs that contain only one vague adjective. Doubly-graded descriptions tend to cause ambiguity, since they involve a trade-off between several dimensions. An NP like 'the tall fat giraffe', for example, might be describe a referent that is neither the tallest nor the fattest giraffe, as long as a combination of height and fatness singles it out. Some of the problems that come up in such cases will be discussed in section 9.1.

### 3. Generation of Crisp Descriptions

Arguably the most fundamental task in Generation of Referring Expressions (GRE), called *Content Determination* (CD) is finding a set of properties which jointly identify the intended referent. Various CD algorithms have been proposed, most of which *approximate* the minimal number of properties that are needed to identify the target. Approximations differ in terms of their computational complexity and the degree to which they match the way in which people use referring expressions (see Dale and Reiter 1995 for a survey). As we shall see in section 8, any one of these algorithms could be used as a basis for our task. For concreteness, we focus here on Dale and Reiter's Incremental Algorithm (IA for short). We shall use a form of the IA that can refer to sets as well as individuals, as long as the sets are individuated via their elements (i.e., distributively, as opposed to collectively, cf., Stone 2000). This version of the IA will be called $\text{IA}_{Plur}$. (For motivation and extensions, see van Deemter 2000, 2002.)

**The Incremental Algorithm.** Put simply, IA accumulates semantic properties until the target objects are the only ones in the domain of which all the accumulated properties are true. This can be done by arranging the properties in a list, and by checking, for each property in the list, whether it is useful (in the sense that it removes one or more distractors); if a property is useful, it is included in the description, after which the next property is given the same treatment. This process of checking and including goes on until the target objects are the only ones of which all the properties in the list are true (i.e., until there are no distractors left).

For reasons that will become apparent later, we complicate matters slightly: following Dale and Reiter, we view each property as consisting of an Attribute (e.g., `colour`) and a Value (e.g., white), written ⟨Attribute,Value⟩. (Attributes can be viewed as grouping

together a number of related properties.) Attributes are ordered in a list $A$, and this *preference order* determines the order in which properties are examined (and possibly added to the description) by the algorithm. Suppose $S$ is the target set, and $C$ the set of all objects that play a role at a given stage of the algorithm (we call these the confusables). The algorithm iterates through $A$; for each Attribute, it checks whether, by specifying a Value for it, one can rule out at least one member of $C$ that has not yet been ruled out; if so then the Attribute is added to a set $L$, with the best possible Value (as determined by `FindBestValue`). Confusables that are ruled out are removed from $C$. The expansion of $L$ and the contraction of $C$ continue until $C = S$:

$\boldsymbol{L} := \oslash$
$\boldsymbol{C} :=$ Domain
For each $\boldsymbol{A_i} \in \boldsymbol{A}$ do

$\boldsymbol{V_i} =$ FindBestValue$(\boldsymbol{S}, \boldsymbol{A_i})$
If $\boldsymbol{S} \subseteq [\![ \langle \boldsymbol{A_i}, \boldsymbol{V_i} \rangle ]\!]$ & $\boldsymbol{C} \not\subseteq [\![ \langle \boldsymbol{A_i}, \boldsymbol{V_i} \rangle ]\!]$ then do

$\boldsymbol{L} := \boldsymbol{L} \cup \{\langle \boldsymbol{A_i}, \boldsymbol{V_i} \rangle\}$
$\boldsymbol{C} := \boldsymbol{C} \cap [\![ \langle \boldsymbol{A_i}, \boldsymbol{V_i} \rangle ]\!]$
If $\boldsymbol{C} = \boldsymbol{S}$ then Return $\boldsymbol{L}$

Return *Failure*

`FindBestValue` selects the 'best value' from amongst the Values of a given Attribute, assuming that these are linearly ordered in terms of specificity. The function selects the Value that removes most distractors, but in case of a tie, the least specific contestant is chosen, as long as it is not less specific than the basic-level Value (i.e., the most commonly occurring and psychologically most fundamental level, Rosch 1987). IA$_{Plur}$ can refer to individuals as well as sets, since reference to a target individual $r$ can be modelled as reference to the singleton set $\{r\}$.

**Existing treatment of gradables.** IA$_{Plur}$ deals with vague properties in essentially the same way as FOG: Attributes like `size` are treated as if they were not context dependent: their Values always apply to the same objects, regardless of what other properties occur in the description. In this way, IA could never describe the same animal as 'the large chihuahua' and 'the small brown dog', for example. This approach does not do justice to gradable adjectives, whether they are used in the base form, the superlative, or the comparative. Suppose, for example, one set a fixed quantitative boundary, making the word 'large' true of everything above it, and false of everything below it. Then IA would tend to have little use for this property at all since, presumably, every chihuahua would be small and every alsacian large, making each of the combinations $\{large, chihuahua\}$ (which denotes the empty set) and $\{large, alsacian\}$ (the set of all alsacians) useless. In other words, existing treatments of gradables in GRE fail to take the 'efficiency of language' into account (Barwise and Perry 1983, see our section 2).

## 4. Generation of Vague Descriptions

We now turn to the question how vague descriptions may be generated from numerical data. We focus on semantic issues, postponing discussion of Pragmatics until section 5, and Linguistic Realisation until section 6. We shall make occasional reference to a PROLOG program called VAGUE, designed by Richard Power, which implements a version of the algorithm descibed in this section. Code and documentation of VAGUE can be found at `http://www.csd.abdn.ac.uk/~kvdeemte/vague.html`.

### 4.1 Expressing one vague property

**Numerical properties.** We shall assume that gradable properties are stored in the Knowledge Base (KB) as Attributes with (decimal) numerical Values, where the numbers can be the result of physical measurements. We will sometimes speak of these numerical Values as if they represented exact Values even though they typically represent

approximations.[3] For concreteness, we shall take them to be of the form $n$ cm, where $n$ is a positive real number. For example,

`type` = *rodent, mouse*
`colour` = *black, blue, yellow*
`size` = 3cm, 4cm,..., 10cm.

Making use of this KB, the IA is able to generate a description involving a list of properties like $L = \{$yellow,mouse,9cm$\}$, for example, exploiting the Attribute `size`. The result could be the NP 'The 9-cm yellow mouse', for example. The challenge formulated in section 1, however, is to avoid unnecessary precision, by avoiding numerical values unless they are necessary for the individuation of the target. This challenge will be answered using a *replacement* strategy. Numerical Values such as 9cm, in $L$, will be replaced by a superlative Value ('being the unique largest element of $C$') whenever all distractors happen to have a smaller size. This list can then be realised in several ways, using either the superlative, the comparative, or the base form (e.g., 'The *largest* yellow mouse', 'The *larger* yellow mouse', or 'The *large* yellow mouse').

**Exploiting numerical properties, singular.** To (almost[4]) ensure that every description contains a property expressible as a noun, we shall assume that the `type` Attribute is more highly preferred than all others. Suppose also, for now, that properties related to size are less preferred than others. As a result, all other properties that turn up in the NP are already in the list $L$ when size is added. Suppose the target is $c_4$:

`type`($c_1$)=`type`($c_2$)=`type`($c_3$)=`type`($c_4$)=mouse
`type`($p_5$)=rat
`size`($c_1$)=6cm
`size`($c_2$)=10cm
`size`($c_3$)=12cm
`size`($c_4$)=`size`($p_5$)=14cm

Since gradable properties are (for now at least) assumed to be dispreferred, the first property that makes it into $L$ is 'mouse', which removes $p_5$ from the context set. (Result: $C = \{c_1, ..., c_4\}$.) Now `size` is taken into account, and `size`$(x) =$ 14cm singles out $c_4$. The resulting list is

$L = \{$mouse, 14cm$\}$

This might be considered the end of the matter, since the target has been singled out. But we are interested in alternative lists, to enable later modules to use gradable adjectives.

---

3 The degree of precision of the measurement (James et al. 1996, section 1.5) determines which objects can be described by the GRE algorithm, since it determines which objects count as having the same size.

4 To turn this likelihood into a certainty, one can add a test at the end of the algorithm, which adds a type-related property if none is present yet (cf., Dale and Reiter 1995). VAGUE uses both of these devices.

One way in which such a list can be computed is as follows. Given that 14cm happens to be the greatest size of any mouse, $\texttt{size}(x) = 14$cm can be replaced, in $L$, by the property of 'being the sole object larger than all other elements of $C$' (notation: $size(x) = max_1$; note that $C$ is the set of mice). Since this property is only applicable because of the properties earlier-introduced into $L$, it becomes essential that $L$ is an ordered list:

$$L = \langle \text{mouse}, \texttt{size}(x) = max_1 \rangle \quad \text{('the largest mouse')}$$

**Exploiting numerical properties, plural.** If plural descriptions were generated using the replacement strategy sketched above, it would be impossible to characterise sets whose elements have different sizes. To make this possible, we have to use inequalities, that is, Values of the form '$> \alpha$' or '$< \alpha$', instead of Values of the form '$= \alpha$'. Therefore, we compile the KB into a more elaborate form by replacing equalities by *in*equalities of the form $\texttt{size}(x) > \alpha$ or $\texttt{size}(x) < \alpha$. The new KB can be limited to *relevant* inequalities only: for every $n$ such that the *old* KB contains an equality of the form $\texttt{size}(x) = n$ cm, the *new* KB contains all those inequalities whose truth follows from the equalities in the old KB. For example,

$\texttt{size}(c_4), \texttt{size}(p_5) > 12$ cm
$\texttt{size}(c_3), \texttt{size}(c_4), \texttt{size}(p_5) > 10$ cm
$\texttt{size}(c_2), \texttt{size}(c_3), \texttt{size}(c_4), \texttt{size}(p_5) > 6$ cm,

where `'size'` is an Attribute, '$> 12$ cm', '$> 10$ cm', and '$> 6$ cm' are Values, and $c_2$, $c_3$, $c_4$, $c_5$, $p_5$ are domain objects of which a given $\langle Attribute, Value \rangle$ combination is true. The procedure is analogous to the treatment of negations and disjunctions in Van Deemter (2002): properties that are implicit in the KB are made available for GRE.

The representation of inequalities is not entirely trivial. For one thing, it is convenient to view properties of the form $\texttt{size}(x) < \alpha$ as belonging to a different Attribute than those of the form $\texttt{size}(x) > \alpha$, because this causes the Values of an Attribute to be linearly ordered: being larger than 12cm implies being larger than 10cm, and so on. More importantly, it will now become normal for an object to have many Values for the same Attribute; $c_4$, for example, has the Values $>$ 6cm, $>$ 10cm, and $>$ 12cm. Each of these Values has equal status, so the notion of a basic-level Value cannot play a role (cf., Dale and Reiter 1995). If we abstract away from the role of basic-level Values, then Dale and Reiter's `FindBestValue` chooses the most general Value that removes the maximal number of distractors, as we have seen. The problem at hand suggests a simpler approach which will always prefer logically stronger inequalities over logically weaker ones, even when they do not remove more distractors.[5] (Thus, $\texttt{size}(x) > m$ is preferred over $\texttt{size}(x) > n$ iff $m > n$; conversely, $\texttt{size}(x) < m$ is preferred over $\texttt{size}(x) < n$ iff $m < n$.) This is reflected by the order in which the properties are listed above: once a `size`-related property is selected, later `size`-related properties do not remove any distractors and will therefore not be included in the description.

Let us return to our example. Suppose the target set $S$ is $\{c_3, c_4\}$. The KB models its two elements as having different sizes (12cm and 14cm, respectively), hence they do not share a property of the form $\texttt{size}(x) = \alpha$. They do, however, share the property $\texttt{size}(x) > 10$cm. This property is exploited by $\text{IA}_{Plur}$ to construct the list

$$L_1 = \langle \text{mouse}, >10cm \rangle,$$

---

[5] A statement $p$ is *logically stronger* than $q$ if $p$ has $q$ as a logical consequence (i.e., $p \models q$) while the reverse is not true (i.e., $q \not\models p$).

first selecting the property 'mouse', then the property $\texttt{size}(x) > 10\text{cm}$. (The property $\texttt{size}(x) > 12\text{cm}$ is attempted first but rejected.) Since $L$ succeeds in distinguishing the two target elements, it follows that they are the *only* mice greater than 10cm. Consequently, this inequality can be replaced by the property 'being a set of cardinality 2, whose elements are larger than all others' (notation: $size(x) = max_2$), leading to NPs such as 'the largest (two) mice':

$$L_2 = \langle \text{mouse}, size(x) = max_2 \rangle.$$

Note that $size(x) = max_2$ is true of a *pair* of mice: strictly speaking, the step from $L_1$ to $L_2$ translates a distributive property ('being larger than 10cm') into a collective one. The case in which the numeral is 1 corresponds with the singular (e.g., 'the largest mouse'). Optionally, we can go a step further and replace $size(x) = max_2$ by the less specified property $size(x) = max$, which abbreviates 'being a set of cardinality greater than 1, all of whose elements are larger than all other elements in $C$'. The result may be realised as 'the largest mice'.

$$L_3 = \langle \text{mouse}, size(x) = max \rangle.$$

**Ordering of properties.** Even if comparative properties are at the bottom of the preference order, while stronger inequalities precede weaker ones, the order is not fixed completely. Suppose, for example, that the KB contains information about `height` as well as `width`, then we have inequalities of the forms (a) $\texttt{height} > x$, (b) $\texttt{height} < x$, (c) $\texttt{width} > x$, and (d) $\texttt{width} < x$. Which of these should come first? Hermann and Deutsch (1976; also reported in Levelt 1989) show that greater *differences* are most likely to be chosen, presumably because they are more striking. In experiments involving candles of different heights and widths, if the referent is both the tallest and the fattest candle, subjects tended to say 'the *tall* candle' when the tallest candle is much taller than all others while the same candle is only slightly wider than the others; if the reverse is the case, the preference switches to 'the *fat* candle'. Hermann and Deutsch's findings may be implemented as follows. Firstly, the Values of the different Attributes should be normalised to make them comparable. Secondly, preference order should be calculated dynamically (i.e., based on the current value of $C$, and taking the target into account), preferring larger gaps over smaller ones. (It is possible, for example, that `width` is most suitable for singling out a *black* cat, but `height` for singling out a *white* cat.) The rest of the algorithm remains unchanged.

**Beyond Content Determination.** Assuming the analysis of section 2.1, 'The $n$ large mouse/mice' is *semantically* equivalent to 'The $n$ largest mouse/mice'. Consequently, there is no need to distinguish between the two at the level of CD. Representations like the ones in $L_2$ and $L_3$ are neutral between the superlative and the base form. Pragmatic constraints determine which of these expressions ('the ($n$) largest, 'the ($n$) larger', 'the ($n$) large') is most appropriate in a given situation. (Section 5.)

**Inference.** The replacement strategy, whereby one list of properties is transformed into another, is essentially a simple kind of logical inference. $L_1$ and $L_2$, for instance, are guaranteed to single out the same set, given that exactly two mice are larger than 10cm; given the content of the KB, the two lists are co-extensive. Once the numeral is dropped, however, as in $L_3$, there is real loss of information: $L_3$ can be used for characterising a number of sets, including the one characterised by $L_2$. In any case, the properties in these lists are logically distinct, so the choice between them belongs to CD.

### 4.2 Expressing several vague properties

If the KB contains several gradable Attributes, a description can make use of several of them (7). Even if only one gradable Attribute is represented, descriptions may contain different adjectives, expressing opposites (8).

7. *The tallest two of the smallest three mice.*
8 *The mice that are taller than* 2*cm but shorter than* 4*cm*.

(The latter may be better expressed as *The mice that are between* 2 *and* 4*cm tall*.) Let us see how the algorithm of the previous sections can be extended to these cases.

**Descriptions using (in)equalities.** When opposites are part of the KB, there is no need for representing equalities separately, since they arise automatically, as combinations of opposites. Every equality of the form '$\texttt{size}(x) = m$ cm' is equivalent to the combination of a property of the form '$\texttt{size}(x) > i$ cm' and one of the form '$\texttt{size}(x) < j$ cm'. Given the content of the following KB, for example, saying that the size of an object is between 6cm and 12cm amounts to saying that its size is 10cm, and this is implemented by adding appropriate transformations to the generator.

$\texttt{size}(c_1) < 10$ cm
$\texttt{size}(c_1)$, $\texttt{size}(c_2) < 12$ cm
$\texttt{size}(c_1)$, $\texttt{size}(c_2)$, $\texttt{size}(c_3) < 14$ cm

$\texttt{size}(c_4)$, $\texttt{size}(p_5) > 12$ cm
$\texttt{size}(c_3)$, $\texttt{size}(c_4)$, $\texttt{size}(p_5) > 10$ cm
$\texttt{size}(c_2)$, $\texttt{size}(c_3)$, $\texttt{size}(c_4)$, $\texttt{size}(p_5) > 6$ cm

Different measures have to be taken when several vague Attributes are involved. Suppose height has these Values:

$\texttt{height}(c_1) = 7$ cm
$\texttt{height}(p_5) = 8$ cm
$\texttt{height}(c_3) = 9$ cm
$\texttt{height}(c_2) = \texttt{height}(c_4) = 10$ cm.

After recompiling these into the form of inequalities (reiterating types):

$\texttt{type}(c_1)=\texttt{type}(c_2)=\texttt{type}(c_3)=\texttt{type}(c_4)$=mouse
$\texttt{type}(p_5)$=rat

$\texttt{height}(c_1) < 8$ cm
$\texttt{height}(c_1)$, $\texttt{height}(p_5) < 9$ cm
$\texttt{height}(c_1)$, $\texttt{height}(c_3)$, $\texttt{height}(p_5) < 10$ cm

$\texttt{height}(c_2)$, $\texttt{height}(c_4) > 9$ cm
$\texttt{height}(c_2)$, $\texttt{height}(c_3)$, $\texttt{height}(c_4) > 8$ cm
$\texttt{height}(c_2)$, $\texttt{height}(c_3)$, $\texttt{height}(c_4)$, $\texttt{height}(p_5) > 7$ cm

Suppose the target set is $\{c_2, c_3\}$. The algorithm will first select the property mouse, since crisp properties are more preferred than vague ones. (Result: $C = \{c_1, c_2, c_3, c_4\}$). The sequel depends on preference order. Omitting the property of being a mouse for brevity, possible results include the following:

*(a)* $L_a = \langle\ \texttt{size} < 14$ cm , $\texttt{height} > 8$ cm $\rangle$, to be realised as, e.g., 'the mice taller than 8cm but smaller than 14cm'.
*(b)* $L_b = \langle\ \texttt{height} > 8$ cm, $\texttt{size} > 6$ cm, $< 14$ cm $\rangle$, e.g., 'the mice that are taller than 8cm and sized between 6 and 14cm'.

Analogous to section 4.1, one might stop here. But there is scope here for logical inference, even more so than before; likewise, there are pitfalls, more than before.

**Adjectives in superlative and base form.** To generate descriptions like the ones in examples (7, 8), we need to transform a comparative property into a superlative property, moving from properties of the form '$\texttt{height} > x$' to properties of the form 'the tallest $n$'. This can be done in different ways. For example, $L_a$ may give rise to

$(i)$ $\langle$ size $< 14$ cm, height$(x) = max_2\rangle$,
('The tallest two of the mice that are smaller than 14cm')
$(ii)$ $\langle$ size$(x) = min_3$, height$(x) = max_2\rangle$
('The tallest two of the smallest three mice')

Once we know which of these outcomes is preferable, the algorithm may be finetuned. (If brevity is an issue, for example, then one might let a generation program vary the preference order used by the IA, then choose the outcome that is shortest.) The transformations described so far rest on logical equivalence (modulo the KB). If numerals are omitted as well, the result is usually no longer equivalent of course, and the description is at risk of becoming almost entirely uninformative (e.g. $L_2$):

$L_1 = \langle$ size$(x) = min_3$, height$(x) = max\rangle$
$L_2 = \langle$ size$(x) = min$, height$(x) = max\rangle$

The algorithm outlined in this and the previous section can be summarised as follows:

**GRE for Vague Descriptions (using IA)**:

1. Construct KB using Attributes and Values, assigning numerical Values to Gradable Attributes.
2. Recompile the KB, replacing equalities by inequalities, for all gradable Attributes.
3. Determine the preference order between the different groups of Attributes. (A safe approach is to give all gradable Attributes lower preference than all non-gradable ones.)
4. Run IA$_{Plur}$ (section 3.2), resulting in a list of properties that jointly identify the target.
5. Apply inferences to the list of properties. For example replace combinations of inequalities by one exact Value; replace inequalities by properties that involve a cardinality; and so on.
6. Perform Linguistic Realisation (section 6).

If gradable properties are less preferred than crisp ones (point 3) then this algorithm will only use gradable properties if an entirely crisp distinguishing description is impossible. This may well cause gradable properties to be under-used. For this and other reasons, we shall consider *non-* incremental versions of these ideas in section 8.

### 4.3 Computational Complexity

We will examine the worst-case complexity of interpretation as well as generation, to shed some light on the hypothesis that vague descriptions are more difficult to process than others, because they involve a comparison between objects (Beun and Cremers 1998, Krahmer and Theune 2002). Before we do this, consider the tractability of the original IA. If the running time of FindBestValue$(r, A_i)$ is a constant times the number of Values of the Attribute $A_i$, then the worst-case running time of IA (and IA$_{Plur}$) is $O(n_v n_a)$, where $n_a$ equals the number of Attributes in the language and $n_v$ the average number of Values of all Attributes. This is because, in the worst case, all Values of all Attributes need to be attempted (Van Deemter 2002). As for the new algorithm, we focus on the crucial phases 2, 4 and 5.

**Phase 2**: Recompilation of the KB forces one to compare all domain elements with each other. This takes at most quadratic time (i.e., $O(n^2)$, where $n$ is the number of elements in the domain). This can be done off line, once and for all.

**Phase 4**: Content Determination. The initial list of properties, which contains inequalities, (e.g., $L = \langle$ mouse, $>$ 5cm $\rangle$) is calculated by IA$_{Plur}$. The algorithm has to take more Attribute/Value pairs into account as a result of the recompilation of the KB), but this does not change its theoretical complexity (using $n_v$ and $n_a$ as variables): it is $O(n_v n_a)$.

**Phase 5**: Inference. The only inference step described so far replaces an inequality (e.g.,

$\texttt{height} > n$cm) by a 'superlative' property (e.g., $\texttt{height} = max_2$). This step requires no computation to speak of: for any given inequality that appears in the description, the value of $m$ can be read off the input to the generator in $O(n_d)$ steps, where $n_d$ is the number of distractors. (This comes down to counting the number of elements in the extension of the inequality.) Therefore, if the number of inequalities in the description is $n_i$ then the complexity is $O(n_d n_i)$.

Thus, the complexity of GRE in the gradable case is determined by three steps: the first is quadratic and can be performed off line; the second has a worst-case running time of $O(n_v n_a)$, and the third one has a worst-case running time of $O(n_d n_i)$. Thus, gradable GRE takes only polynomial time, and if we focus on the part that cannot be done off line, it takes only linear time. In other words, gradable GRE does take more time than non-gradable GRE, but the difference seems modest.

The intuition that vague descriptions are more difficult than others is also confirmed (though again only to a modest extent) when we focus on the hearer. First, consider a *non-vague* description consisting of a combination of $n$ properties, $P_1, ..., P_n$. To discover its referent, the denotation of the Boolean expression $P_1 \cap .. \cap P_n$ needs to be calculated, which takes just $n - 1$ calculations of the form

> Intersect $\|P_1\| \cap ... \cap \|P_{i-1}\|$ (a set which has been computed already)
> with $\|P_i\|$ (the extension of the next property in the description).

If computing the intersection of two sets takes constant time then this makes the complexity of interpreting non-vague descriptions linear: $O(n_d)$, where $n_d$ is the number of properties used. In a *vague* description, the property last added to the description is context-dependent. Worst case, calculating the set corresponding with such a property, of the form $\texttt{size}(x) = max_m$ for example, involves sorting the distractors as to their size, which may amount to $O(n_d^2)$ or $O(n_d\ log\ n_d)$ calculations (depending on the sorting algorithm, (Aho et al. 1983). Once again, the most time-consuming part of the calculation can be performed off line, since it is the same for all referring expressions.

Thus, the worst-case time complexity of *interpretation* is as follows: the part that can be computed off line takes $O(n_d\ log\ n_d)$ calculations. The part that has to be computed for each referring expression separately takes $O(n_d)$ calculations. Once again, there is a difference with the non-gradable case, but the difference is modest, especially regarding the part that cannot be done off line. One should bear in mind that worst-case theoretical complexity is not always a good measure of the time that a program takes in the kinds of cases that occur most commonly, let alone the difficulty for a person. For example, it seems likely that hearers and speakers will have most difficulty dealing with differences that are too small to be obvious (e.g., two mice that are very similar in size).

## 5. Pragmatic Constraints

NLG has to do more than select a distinguishing description (i.e., one that unambiguously denotes its referent, Dale 1989): the selected expression should also be felicitous. Consider the question, discussed in the philosophical logic literature, whether it is legitimate, for a gradable adjective, to distinguish between 'observationally indifferent' entities: Suppose two objects $x$ and $y$ are so similar that it is impossible to distinguish their sizes; can it ever be reasonable to say that $x$ is large and $y$ is not? A positive answer would not be psychologically plausible, since $x$ and $y$ are indistinguishable; but a negative answer would prohibit *any* binary distinction between objects that are large and objects that are not, given that one can always construct objects $x$ and $y$ one of which falls just below the divide while the other falls just above it. This is the strongest version of the *sorites* paradox (e.g., Hyde 2002).

Our approach to vague descriptions allows a subtle response: that the offending statement may be correct yet infelicitous. This shifts the problem from asking when vague descriptions are 'correct' to the question when they are used felicitously. Felicity is naturally thought of as a gradable concept. There is therefore no need for a generator to demarcate precisely between felicitous and infelicitous expressions, as long as all the utterances generated are felicitous enough. When in doubt, a generator should avoid the expression in question. If $x$ and $y$ are mice of sizes 10cm and 9.9cm, for example, then it is probably better to describe $x$ as *'the largest mouse'* than as *'the large mouse'*.

Prior to carrying out the experiments to be reported in section 7, we believed that the following constraints should be taken into account:

> **Small Gaps.** Expressions of the form 'The ($n$) large [N]' are infelicitous when the gap between (1) the *smallest* element of the designated set $S$ (henceforth, $s^-$) and (2) the *largest* N smaller than all elements of $S$ (henceforth, $s^+$) is small in comparison with the other gaps (Thorisson 1994, Funakoshi et al. 2004). If this gap is so small as to make the difference between the sizes of $s^-$ and $s^+$ impossible to perceive, then the expression is also infelicitous.
>
> **Dichotomy.** When separating one single referent from one distractor, the comparative form is often said to be favoured ('Use the comparative form to compare two things'). We expected this to generalise to situations where all the referents are of one size, and all the distractors of another.
>
> **Minimality.** Unless *Small Gaps* and *Dichotomy* forbid it, we expected that preference should be given to the base form. In English, where the base form is morphologically simpler than the other two, this rule could be argued to follow from Gricean principles (Grice 1975).

To keep matters simple, Linguistic Realisation could choose the base form if and only if the gap between $s^-$ and $s^+$ surpasses a certain value, which is specified interactively by the user. (This approach was chosen for the VAGUE program.)

As for the presence/absence of the *numeral* in the description, there appear to be different 'believable' patterns of linguistic behaviour. A cautious generator might only omit the numeral when the pragmatic principles happen to enforce a specific extension (e.g., 'the large mice', when the mice are sized 3cm, 2.8cm, 2.499cm, and 2.498cm). This would allow the generator to use vague expressions, but only where they result in a description that is itself unambiguous.

We shall see in section 7 that it has not been easy to confirm the pragmatic constraints of the present section experimentally.

**6. Linguistic Realisation**

Some recent GRE algorithms have done away with the separation between Content Determination and Linguistic Realisation, *interleaving* the two processes instead (Stone and Webber 1998, Krahmer and Theune 2002). We have separated the two phases because, in the case of vague descriptions, interleaving would tend to be difficult. Consider, for instance, the list of properties $L = \langle$ `size` $> 3$ cm, `size` $< 9$ cm $\rangle$. If interleaving forced us to realise the two properties in $L$ one by one, then it would no longer be possible to combine them into, for example, 'the largest mouse but one' (if the facts in the KB support it), or even into 'the mice between 3 and 9cm' (since `size` $> 3$ cm is realised before `size` $< 9$ cm). Clearly, sophisticated use of gradable adjectives requires a separation between CD and Linguistic Realisation, unless one is willing to complicate linguistic realisation considerably.

Having said this, the distinction between CD and Linguistic Realisation is not always easy to draw. We propose to think of it as separating the language-independent, *logical* aspect of referring expressions generation from its language-dependent, *linguistic* aspect. Our algorithm suggests a distinction into three phases, the first two of which can be thought of as part of CD:

> **(1)** CD *proper*, that is, the production of a distinguishing list of properties $L$;
> **(2)** An *inference* phase, during which the list $L$ is transformed;
> **(3)** A *realisation* phase, during which the choice between base, superlative and comparative forms is made, among other things.

One area of current interest concerns the left-to-right arrangement of pre-modifying adjectives within an NP (e.g., Shaw and Hatzivassiloglou 1999, Malouf 2000). Work in this area is often based on assigning adjectives to a small number of categories (e.g., Precentral, Central, Postcentral, and Pre-head) which predict adjectives' relative position. Interestingly, vague properties tend to be realised before others. Greenbaum et al. (1985), for example, report that 'adjectives denoting `size`, `length`, and `height` normally precede other nonderived adjectives' (e.g., 'the small round table' is usually preferred to 'the round small table').

Semantically, this does not come as a surprise. In a noun phrase of the form 'the three small(-est) [N]', for example, the words preceding N select the three smallest elements of [N]. It follows that, to denote the three smallest elements of the set of round tables, the only option is to say 'the three small round tables', rather than 'the three round small tables'. The latter would mean something else, namely 'the three round ones among the $n$ small(est) tables' (where $n$ is not specified). It actually seems quite possible to say this, but only when some set of small tables is contextually salient (e.g., 'I don't mean *those* small tables, I mean the three *round* ones'). Given that $n$ is unspecified, the noun phrase would tend to be very unclear in any other context.

The VAGUE program follows Greenbaum's rule by realising gradable properties before non-gradable ones, choosing some simple (and sometimes stilted) syntactic patterns.

## 7. Empirical grounding

A full validation of a GRE program that generates vague descriptions would address the following questions: (1) When is it natural to generate a vague description (i.e., a qualitative description as opposed to a purely quantitative one)? (2) Given that a vague description is used, which form of the description is most natural? and (3) Are the generated descriptions properly understood by hearers and readers? Much is unknown, but we shall summarise the available results in these three areas very briefly, referring readers to the literature for details.

### 7.1 Human speakers' use of vague descriptions

Common sense (as well as the Gricean maximes, Grice 1975) suggests that vague descriptions are preferred by speakers over quantitative ones whenever the additional information provided by a quantitative description is irrelevant to the purpose of the communication. We are not aware of any empirical validation of this idea, but the fact that vague descriptions are frequent is fairly well documented. Dale and Reiter, for example, discussed the transcripts of a dialogue between people who assemble a piece of garden furniture (originally recorded by Candy Sidner). They found that, while instructional texts tended to use numerical descriptions like 'the $3\frac{1}{4}$" bolt', human assemblers *'unless they were reading or discussing the written instructions, in all cases used*

*relative modifiers, such as* the long bolt' (Dale and Reiter 1995).[6]

Our own experiments (van Deemter 2004) point in the same direction. In one experiment, for example, 34 students at the University of Brighton were shown six pieces of paper, each of which showed two isosceles and approximately equilateral triangles. Triangles of three sizes were shown, with bases of 5mm, 8mm, and 16mm respectively. On each sheet, one of the two triangles had been circled with a pencil. We asked subjects to imagine themselves on the phone to someone who held a copy of the same sheet, but not necessarily with the same orientation (e.g., possibly upside down), and to complete the answers on the dots:

```
Q: Which triangle on this sheet
   was circled?
A: The ............ triangle.
```

This setup was used for testing a number of hypotheses. What is relevant for current purposes is that all except one subject used qualitative size-related descriptions ('the big triangle', 'the largest triangle', etc.) in the vast majority of cases. As many as 27 of the 34 subjects used such descriptions in *all* cases.

It seems likely that qualitative descriptions would be less frequent if speakers were offered an easy way to determine the relevant measurements (e.g., if a spring rule was provided). As it was, subjects went for the easy option, relying on a comparison of sizes rather than on an estimation of their absolute values. Further experiments are needed before we can say with more confidence under what circumstances vague descriptions are favoured over absolute ones.

It is normally perhaps unlikely that people produce language on the basis on the kind of numerical representations that our algorithm has used as input. Although psychological plausibility is not our aim, it is worth noting that the *in*equalities computed as step 2 of the algorithm of section 4 might be psychologically more plausible, since they are essentially no more than comparisons between objects.

**7.2 Testing the correctness of the generated expressions**

Sedivy et al. (1999) asked subjects to identify the target of a vague description in a visual scene. Consider 'the tall cup'. The relevant scene would contain three distractors: (1) a less tall object of the same type as the target (e.g., a cup that is less tall), (2) a different kind of object which previous studies had shown to be intermediate in height (e.g., a pitcher that, while being taller than both cups, was neither short nor tall for a pitcher), and (3) a different type of object to which the adjective is inapplicable (e.g., a door key). Across the different conditions under which the experiment was done (e.g., allowing subjects to study the domain before or after the onset of speech), it was found not to matter much whether the adjective applied 'intrinsically' to the target object (i.e., whether the target was tall for a cup): hearers identifed the target without problems in both types of situations. The time subjects took before looking at the target for the first time was measured, and although these latency times were somewhat greater when the referent were not intrinsically tall than when they were, the average difference was tiny at 554 versus 538 miliseconds. Since latency times are thought to be sensitive to most of

---

6 Presumably, Beun and Cremers (1998) found vague adjectives to be rare because, in their experiments, referents could always be identified using non-gradable dimensions.

the problems that hearers may have in processing a text, these results suggest that, for dimensional adjectives, it is forgivable to disregard global context.

To get an idea whether our *plural* descriptions are understood correctly by human readers, we showed subjects sequences of numbers, exactly two of which appeared in brackets, along with the following instructions:

```
Suppose you want to inform a hearer *which numbers in a given
list appear in brackets*, where the hearer knows what the numbers
are, but not which of them appear in brackets. For example, the
hearer knows that the list is

            1   2   1   7   7   1   1   3   1

You, as a speaker, know that only the two occurrences of the
number 7 appear in brackets:

            1   2   1  (7) (7)  1   1   3   1

Our question to you is: Would it be *correct* to convey this
information by saying ''The two high numbers appear in brackets''?
```

The outcomes of the experiment suggested that readers understand plural vague descriptions in accordance with the semantics of section 2 (van Deemter 2000). In other words, they judged the description to be correct if and only if the two highest numbers in the sequence appeared in brackets.

Assessing the evidence, it seems that vague descriptions are largely unproblematic from the point of view of interpretation.

**7.3 Testing the felicity of the generated expressions**

How can we choose between the different forms that a vague description can take? Reiter and Sripada (2002) showed that the variation in corpora based on expert authors can be considerable, especially in their use of vague expressions (e.g., 'by evening', 'by late evening', 'around midnight'). We confirmed these findings using experiments with human subjects (van Deemter 2004), focussing on the choice between the different forms of the adjective. Informally:

1. The Dychotomy constraint of section 5 did not hold up well: Even when comparing two things, the superlative form was often preferred over the comparative.
2. When base forms were used, the gap was almost invariably large.
3. Yet, the Minimality constraint of section 5 turned out to be difficult to confirm: Even when the gap was large, base forms were often dispreferred.

The validity of these results can be debated (van Deemter 2004) but, taking them at face value, one could base different generation strategies on them. For example, one might use the superlative all the time, since this was – surprisingly – the most frequent form overall. Based on point (2), however, one might also defend using the base form whenever the gap is large enough (as was done in the VAGUE program). Future experiments should allow us to refine this position, perhaps depending on factors such as genre, communicative goal, and type of audience.

**8. Incrementality: help or hindrance?**

The account sketched in section 4 was superimposed on an incremental GRE algorithm, partly because incrementality is well established in this area (Appelt 1985, Dale and Re-

iter 1995). But IA may be replaced by any other reasonable[7] GRE algorithm, for example one that always exactly minimises the number of properties expressed, or one that always 'greedily' selects the property that removes the maximum number of distractors. Let $G$ be any such GRE algorithm, then we can proceed as follows:

**GRE for Vague Descriptions (version not relying on IA)**:

1. Construct KB using Attributes and Values, assigning numerical Values to gradable Attributes.
2. Recompile the KB, replacing equalities by inequalities.
3. Let $G$ deliver an unordered set of properties which jointly distinguish the target if such a set exists. (One or more of these properties may be inequalities.)
4. Impose a linear ordering on the properties produced by (3). (If one wishes to generate the same descriptions as in sections 4.1 and 4.2, then inequalities go last.) Delete any inequalities that do not remove any distractors.
5. Apply inferences (in the style of section 4.1) to the list of properties.
6. Perform Linguistic Realisation.

Imposing a linear order (4) is a necessary preparation for (5) because the superlative properties resulting from (5), unlike the inequalities resulting from (4), are context dependent. For example, $\langle$mouse, $size(x) = max_2\rangle$ (the largest two mice, $\{c_3, c_4\}$) does not equal $\langle size(x) = max_2$, mouse $\rangle$ (the mouse among the largest two elements, $\{c_4\}$). Deletion of superfluous inequalities avoids saying, for example, 'the short(est) black mouse' if there is only one black mouse, because this might invite false implicatures.

**Problems with incrementality.** While IA is generally thought to be consistent with findings on human language production (Hermann and Deutsch 1976, Levelt 1989, Pechmann 1989, Sonnenschein 1982), the hypothesis that incrementality is a good model of human GRE seems unfalsifiable until a preference order is specified for the properties on which it operates. (Wildly redundant descriptions can result if the 'wrong' preference order are chosen.) We shall see that vague descriptions pose particular challenges to incrementality.

One question emerges when the IA is combined with findings on word order and incremental *interpretation*. If human speakers and/or writers perform CD incrementally, then why are properties not expressed in the same order in which they were selected? This question is especially pertinent in the case of vague expressions, since gradable properties are selected last, but realised first (section 6). This means that the Linguistic Realisation cannot start until CD is concluded, contradicting eye-tracking experiments suggesting that speakers start speaking while still scanning distractors (Pechmann 1989). A similar problem is discussed in the psycholinguistics of *interpretation* (Sedivy et al. 1999): interpretation is widely assumed to proceed incrementally, but vague descriptions resist strict incrementality, since an adjective in a vague description can only be fully interpreted when its comparison set is known. Sedivy and colleagues resolve this

---

[7] Concretely, we require of a reasonable GRE algorithm that it avoid combining logically *comparable* inequalities, such as $\texttt{size}(x) > 10$ and $\texttt{size}(x) > 20$, inside one description. All GRE algorithms that we know of fulfill this requirement.

quandary by allowing a kind of revision, whereby later words allow hearers to refine their interpretation of gradable adjectives. – Summarising the situation in generation and interpretation, it is clear that the last word on incrementality has not been said.

**Low preference for gradable properties?** It has been argued that, in an incremental approach, gradable properties should be given a low preference ranking because they are difficult to process (Krahmer and Theune 2002). We have seen in section 4.3 that generation and interpretation of vague descriptions does have a slightly higher computational complexity than that of non-vague descriptions. Yet, by giving gradable properties a low ranking, we might cause the algorithm to under-use them, for example in situations where gradable properties are highly relevant to the purpose of the discourse (e.g., a fist fight between people of very different sizes). Luckily, there are no semantic or algorithmic reasons for giving gradables a low ranking. Let us see how things would work if they were ranked more highly.

Suppose comparative properties do *not* go to the end of the preference list. After transformation into superlative properties, this alternative preference ranking could lead to a list like $\langle$mouse, $size(x) = min_4$, brown, $weight(x) = max_2\rangle$, where two ordinary properties are separated by a superlative one. A direct approach to Realisation might word this as 'The two heaviest brown ones among the smallest four mice'. To avoid such awkward expressions, one can change the order of properties *after* CD (mirroring step (4) above), moving the inequalities to the end of the list before they are transformed into the appropriate superlatives. The effect would be to boost the number of occurrences of gradable properties in generated descriptions while keeping CD incremental.

## 9. Extensions of the approach

**Relational descriptions**. Some generalisations of our method are fairly straightforward. For example, consider a relational description (cf., Dale and Haddock 1991) involving a gradable adjective, as in *'the dog in the large shed'*. CD for this type of descriptions along the lines of section 4 is not difficult once relational descriptions are integrated with a standard GRE algorithm (Krahmer and Theune 2002, section 8.6.2): suppose an initial description is generated describing the set of all those dogs that are in sheds *over a given size* (say, size 5); if this description happens to distinguishing an individual dog then this legitimises the use of the noun phrase *'the dog in the large shed'*. Note that this is felicitous even if the shed is not the largest one in the domain, as is true for $d_2$ in the following situation: (`contains`-$a$=$b$ means that $a$ is contained by $b$)

`type`($d_1$)=`type`($d_2$)=dog
`type`($c$)=cat
`type`($s_1$)=`type`($s_2$)=`type`($s_3$)=shed
`size`($d_1$)=`size`($d_2$)=`size`($c$)=1m
`size`($s_1$)=3m
`size`($s_2$)=5m
`size`($s_3$)=6m
`contains`-$d_1$=$s_1$
`contains`-$d_2$=$s_2$
`contains`-$c$=$s_3$

In other words, 'the dog in the large shed' denotes 'the dog such that there is no other shed that is equally large or larger *and that contains a dog*'. Note that it would be odd, in the above-sketched situation, to say 'the dog in the larg*est* shed'.

**Boolean combinations.** Generalisations to complex Boolean descriptions involving negation and disjunction (van Deemter 2004) appear to be largely straightforward, except for issues to do with opposites and markedness. For example, the generator will have to decide whether to say 'the patients that are old' or 'the patients that are not

young'.

**9.1 Multidimensionality**

**Combinations of adjectives.** When objects are compared in terms of several dimensions, these dimensions can be weighed in different ways (e.g., Rasmusen 1989). Let us focus on references to an individual referent $r$, starting with a description that contains more than one gradable adjective. The NP 'the tall fat giraffe', for example, can safely refer to an element $b$ in a situation like the one below, where $b$ is the only element that exceeds all distractors with respect to some dimension (a different one for $a$ than for $c$, as it happens) while not being exceeded by any distractors in any dimension:

$\texttt{height}(a) = 5\ m$
$\texttt{height}(b) = \texttt{height}(c) = 15\ m$
$\texttt{width}(a) = \texttt{width}(b) = 3\ m$
$\texttt{width}(c) = 2\ m$

Cases like this would be covered if the decision-theoretic property of Pareto Optimality (e.g., Feldman 1980) was used as the sole criterion: Formally, an object $r \in C$ has a Pareto-Optimal combination of Values $\mathcal{V}$ iff there is no other $x \in C$ such that

(1) $\exists V_i \in \mathcal{V} : V_i(x) > V_i(r)$ and
(2) $\neg\exists V_j \in \mathcal{V} : V_j(x) < V_j(r)$

In our example, $b$ is the only object that has a Pareto-Optimal combination of Values, predicting correctly that $b$ can be called 'the tall fat giraffe'. It seems likely, however, that people use doubly-graded descriptions more liberally. For example, if the example is modified by letting $\texttt{width}(a) = 3.1\ m$, making $a$ slightly fatter than $b$, then $b$ might still be the only reasonable referent of 'the tall fat giraffe'. Many alternative strategies are possible. The *Nash arbitration plan*, for example, would allow a doubly-graded description whenever the product of the Values for the referent $r$ exceeds that of all distractors (Nash 1950; cf., Gorniak and Roy 2003, Thorisson 2004, for other plans).

**Multidimensional adjectives (and colour).** Multidimensionality can also slip in through the backdoor. Consider *big*, for example, when applied to 3D shapes. If there exists a formula for mapping three dimensions into one (e.g., length.width.height) then the result is *one* dimension (`overall-size`), and the algorithm of section 4 can be applied *verbatim*. But if 'big' is applied to a person then it is far from clear that there is one canonical formula for mapping the different dimensions of your body into one overall dimension, and this complicates the situation. Similar things hold for such multi-faceted properties like intelligence (Kamp 1975).

*Colour* terms are a case apart. If colour is modelled in terms of saturation, hue, and luminosity, for instance, then an object $a$ may be classified as 'greener' than $b$ on one dimension (e.g., saturation), but 'less green' than $b$ on another (e.g., hue). This would considerably complicate the application of our algorithm to colour terms, which is otherwise mostly straighforward (section 9.3). ('The green chair', said in the presence of two green-ish chairs, would refer to the one that is closest to prototypical green.) A further complication is that different speakers can regard very different values as prototypical, making it difficult to assess which of two objects is 'greener' even on one dimension (Berlin and Kay 1969, p.10-12). (Ideally, GRE should also take into account that the meaning of colour words can differ across different types of referent. Red as in 'red hair', for example, differs from red as in 'red chair'.)

Different attitudes towards multidimensionality are possible. One possibility is to be cautious and to keep aiming for distinguishing descriptions in the strict sense. In this case, the program should limit the use of vague descriptions to situations where there exists a referent that has a Pareto-optimal combination of Values. Alternatively, one

could allow referring expressions to be ambiguous. It would be consistent with this attitude, for example, to map multiple dimensions into one over-all dimension, perhaps by borrowing from principles applied in *perceptual grouping*, where different perceptual dimensions are mapped into one (e.g., Thorisson 1994). The empirical basis of this line of work, however, is still somewhat weak, so the risk of referential unclarity looms large. Also, this attitude would go against the spirit of GRE, where referring expressions have always been assumed to be *distinguishing*.

**9.2 Salience as a gradable property**

We shall see that a natural treatment of salience falls automatically out of our treatment of vague descriptions. As we shall see, this will allow us to simplify the structure of GRE algorithms, and it will explain why many definite descriptions that look as if they were distinguishing descriptions are actually ambiguous.

**A new perspective on salience.** Krahmer and Theune (2002) have argued that Dale and Reiter's dichotomy between salient and non-salient objects (where the objects in the domain are the salient ones) should be replaced by an account that takes *degrees* of salience into account: No object can be too unsalient to be referred to, as long as the right properties are available. In effect, this proposal (which measured salience numerically) analyses 'the black mouse' as denoting the unique *most salient* object in the domain that is both black and a mouse. Now suppose we let GRE treat salience just like other gradable Attributes. Suppose there are ten mice, five of which are black, whose degrees of salience are 1, 1, 3, 4, and 5 (the last one being most salient), while the other objects in the domain (cats, white mice) all have a higher salience. Then our algorithm might generate this list of properties:

$$L = \langle \text{ mouse, black, salience} > 4 \rangle.$$

This is a distinguishing description of the black mouse whose salience is 5: 'the most salient black mouse'. The simpler description 'the black mouse' can be derived by stipulating that the property of being most salient can be left implicit in English. The salience Attribute *has* to be taken into account by CD, however, and this can be ensured in various ways. For example, instead of testing whether $C \cap [\![ \langle A_i, V_i \rangle ]\!] = \{r\}$, one tests whether $r$ is the most salient element of $C \cap [\![ \langle A_i, V_i \rangle ]\!]$. Alternatively, the algorithm might proceed as usual, performing the usual test (involving $C \cap [\![ \langle A_i, V_i \rangle ]\!] = \{r\}$) but starting with a reduced domain, consisting of the things that are at least as salient as the target $r$: $Domain := \{x \in \text{Domain: } salience(x) \geq r\}$. The two approaches are equivalent in many situations.

**Salience + plurality = ambiguity.** It is now easy to see why plural descriptions are often ambiguous. Taking salience into account as suggested above, the singular 'the black mouse' can only refer to the most salient mouse. But 'the mice' can refer to the most salient two (sized 5 and 4), the most salient three (sized 5, 4 and 3), or to all of them. To disambiguate the description, something like a number can be used (e.g., 'the two mice'), just like in the case of vague descriptions.

When salience is combined with other gradable notions, the likelihood of unclarity is even greater. Consider 'the large(st) dog'. Our analysis predicts ambiguity when size and salience do not go hand in hand.

`Type:` *d1* (dog), *d2* (dog), *d3* (dog), *d4* (dog), *c5* (cat)
`Size:` *d1* (20cm), *d2* (50cm), *d3* (70cm), *d4* (60cm), *c5* (50cm)
`Salience:` *d1* (6), *d2* (4), *d3* (3), *d4* (5), *c5* (6).

If we are interested in the *three* most salient dogs ($d_1$, $d_2$ and $d_4$) then 'the large(est) dog' designates *d4*, but if we are interested in the *four* most salient ones ($d_1$, $d_2$, $d_3$ and $d_4$), then it designates $d_3$, for example. In other words, the description is ambiguous between

*d3* and *d4*, depending on whether we attach greater importance to salience or size. This is borne out by our generation algorithm. Consider the simpler of the two treatments of salience, for example, which starts out with a reduced domain. If $d_4$ is the target then the reduced domain (consisting of all things at least as salient as the target) is $\{d_1, d_2, d_4, c_5\}$; 'dog' narrows this down to $\{d_1, d_2, d_4\}$, after which 'size $= max_1$' generates 'the large dog'. But if $d_3$ is the target then the same procedure applies, this time starting with the full domain (since no element is less salient that $d_3$) and the same description is generated to refer to a different animal. For a reader, clearly, salience and gradable adjectives are a problematic combination. This should come as no surprise, since salience itself is a gradable property, and combinations of gradable properties are always problematic, as we have seen in the previous section.

**Salience as a multidimensional property.** Note that salience itself is multidimensional. Consider two people talking about 'the railway station', when one railway station is near but of only minor importance (e.g., only few trains stop there), while another is further afield but of greater significance for travel. In such a situation, it can be unclear which of the two railway stations is intended. Without more empirical research, we cannot know how people combine salience with other dimensions.

GRE has usually assumed that distinguishing descriptions are the norm, but once salience is taken into account (especially in combination with plurals and/or other gradable dimensions) it becomes difficult to generate descriptions that are immune to being misunderstood.

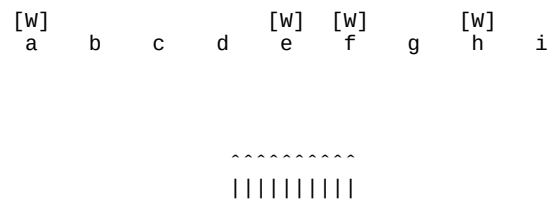
**9.3 Beyond Vague Descriptions: nouns, and pointing**

**Nouns.** Two other generalisations are worth mentioning. The first involves a class of descriptions that do not involve any overt gradable adjectives. Colour terms, for example (cf., section 9.1), are applicable to different degrees, and the same is true for many other nouns, such as 'girl', which involves a vaguely defined age. Similar claims can be made about less obvious cases. Consider a gathering containing one famous professor (a), one junior lecturer (b), one PhD student (c), and a policeman (e). Then the word 'academic' might denote (a), but also (b) or (c). Accordingly, each of the following referring expressions appears viable, mirroring examples 3-4 of section 2:

1. *the academic* (Can only refer to a)
2. *both academics* (Can only refer to $\{a, b\}$)
3. *the three academics* (Can only refer to $\{a, b, c\}$)

These descriptions are easily generated on the basis of a KB that involves Values representing *degrees* of being an academic, the more so because our approach generalises to *ordinal* measurements (except for Small Gaps (section 5), which requires an interval or ratio scale, since it involves an assessment of the size of the gap between Values). Note that this treatment could cover all those nouns that are used with various degrees of strictness. It is difficult to say how many nouns fall in this category, but the phenomenon appears to be widespread. An uncomfortable consequence of these observations is that it is no longer obvious which words denote a crisp property, and which a gradable property. (For example, it is not clear whether GRE should treat 'academic' as gradable.)

**Pointing.** To show that vagueness is also inherent in multimodal communication, imagine the same gathering, but with some more people present. Suppose someone *points* at the centre of the gathering. (See below, where `W` denotes women.) If the distance between pointer and pointee is considerable then the boundaries of the region pointed to are not exactly defined: $e$ is definitely pointed at, but $d$ and $f$ might be doubtful:




Here, $\{e\}$ is a possible referent, and so are $\{d, e, f\}$ and perhaps $\{c, d, e, f, g\}$. The set $\{d, f\}$ is not, since there is a gap between its two elements. If *precise* pointing is represented as a crisp property whose denotation equals the set of elements pointed at (Krahmer and Van der Sluis 2003), then *vague* pointing can be incorporated in our algorithm by representing it by a *gradable* property: we let the generator use a KB that involves numerical *degrees* of being pointed at, where this degree is highest for $e$, next highest for $d$ and $f$, and so on. If this is done, the generator can generate 'these two women', along with a pointing gesture like the one in our example, to refer to $\{e, f\}$. No changes to the algorithm of section 4 are necessary. A variant of this approach arises if pointing is modelled as a way of establishing degrees of *salience* (i.e., the closer to the center of pointing, the higher the value for the attribute SALIENCE) in the style of section 9.2.

## 10. Conclusion

If the usefulness of NLG resides in its ability to present data into human-accessible form, then vagueness must surely be one of its central instruments, because it allows the suppression of irrelevant detail. In principle, this might be done by providing the generator with vague input – in which case no special algorithms are needed – but suitably contextualised vague input is often not available Mellish (2000). The only practical alternative is to provide the generator with 'crisp' (i.e., quantitative) input, allowing the generator to be hooked on to a general-purpose database. It is this avenue that we have explored in this paper, in combination with various (incremental and other) approaches to GRE.

Far from being a peculiarity of a few adjectives, vagueness is widespread. We believe that our approach can be applied to a variety of situations in which vagueness affects referring expressions including, for example,

- Colour terms (section 9.1).
- Nouns that allow different degrees of strictness (section 9.3).
- Degrees of salience (section 9.2).
- Imprecise pointing (section 9.3).

On the other hand, we have also met some considerable obstacles on our way:

**Expressive choice (sections 4 and 7)**. By enabling the generator to produce more referring expressions, we have made it harder to choose between them. For example, when is a qualitative description preferable over a quantitative one? At a more detailed level, the generator must choose between descriptions like 'The heaviest two of the smallest three mice', 'The mice that weigh between 40 and 60 grams', and so on, each of which may single out the same individuals. Section 7.3 has summarised some experimental evidence related to such choices, focussing on the different forms of the adjective, but the evidence is far from conclusive. Much is still unknown, differences between speakers abound, and the experimental methodology for advancing the state of the art in this area is not without its problems (van Deemter 2004).

**Architecture (section 6)**. The inference rules that were necessary to convert one list of properties into another do not sit comfortably within the received NLG pipeline (e.g.,

Reiter and Dale 2000). An example of such an inference rule is the one that transforms a list of the form $\langle$mouse, $>10cm\rangle$ into one of the form $\langle$mouse, $size(x) = max_2\rangle$ if only two mice are larger than 10cm. The same issues also make it difficult to interleave CD and Linguistic Realisation as proposed by various authors, because properties may need to be combined before they are expressed.

**Incrementality (section 8)**. Gradable adjectives complicate the notion of incrementality, in generation as well as interpretation. Focussing on generation, for example, they force us to re-examine the idea that properties can be put into words more or less as soon as they have been selected by Content Determination (even apart from the issue noted under Architecture).

**Adjectives and presuppositions (section 2)**. Our generation-oriented perspective sheds some doubt on Bierwisch's claim that dimensional adjectives are insensitive to standards provided by the global context: If a man's height is 205cm, then surely no local context can make it felicitous (as opposed to just humourous) to refer to him as 'the short man'. A related issue that we have not touched upon is the fact that adjectives are often used partly to pass judgement: One and the same car might be designed as 'the expensive car' by a hesitant customer and as 'the luxury car' by an eager salesman: even if expense and luxury go hand in hand, the two adjectives have different connotations, and this is something that a generator would ideally be aware of.

**Multidimensionality (section 9.1)**. We know roughly how to deal with *one* gradable dimension: 'The short man', for example, is the shortest man around. But in practice, we often juggle several dimensions. This happens, for example, when two adjectives are used ('the short thin man'), or when salience is taken into account (e.g., 'the short man', when the shortest man is not the most salient one), threatening to make irrefutably distinguishing descriptions something of an exception. (For a study of approaches to multidimensionality in a different area, see Masthoff 2004.) At some point, GRE may have to abandon the strategy of aiming for unambiguous descriptions in all situations.

## Acknowledgments

I thank Hua Cheng, Roger Evans, Albert Gatt, Markus Guhe, Imtiaz Khan, Emiel Krahmer, Judith Masthoff, Chris Mellish, Oystein Nilsen, Manfred Pinkal, Paul Piwek, Ehud Reiter, Graeme Ritchie, Ielka van der Sluis, Rosemary Stevenson (†), Matthew Stone, and Sebastian Varges, for helpful comments. I am especially grateful to Richard Power, for inspiration as well as for implementing the VAGUE program at great speed. Thanks are due to four anonymous reviewers for some very substantial contributions. This work has been supported by the EPSRC (GR/S13330, TUNA project).

**11. References**

- Appelt 1985. Doug Appelt. Planning English referring expressions. Artificial Intelligence, 26: 1-33. Reprinted in: B. J. Grosz, K. Sparck Jones, and B. L. Webber (Eds.) (1986). Readings in Natural Language Processing. Los Altos, Ca.: Morgan Kaufmann.

- Barwise and Perry 1983. Jon Barwise and John Perry. *Situations and Attitudes.* MIT Press, Cambridge (Mass.) and London.

- Berlin and Kay 1969. Brent Berlin and Paul Kay. *Basic Color Terms*. Univ. of California Press, Berkeley and Los Angeles.

- Berry et al. 2002. Dianne C. Berry, Peter R. Knapp, and Theo Raynor. Is 15 percent very common? Informing people about the risks of medication side effects. *International Journal of Pharmacy Practice* **10**: 145-151.

- Beun and Cremers 1998. Robbert-Jan Beun and Anita Cremers. Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition* **6**(1/2): 121-152.

- Bierwisch 1989. Manfred Bierwisch. The semantics of gradation. In M.Bierwisch and E.Lang (Eds.) *Dimensional Adjectives*. Berlin, Springer Verlag, 71-261.

- Dale 1989. Robert Dale. Cooking up referring expressions. In Procs. of 27th annual meeting of Accorc. for Comp. Ling. (ACL-89), pages 68-75.

- Dale and Haddock 1991. Robbert Dale and Nickolas Haddock. Generating referring expressions containing relations. Procs. of the 5th Conference of the European Chapter of the ACL, EACL-91.

- Dale and Reiter 1995. Robbert Dale and Ehud Reiter. Computational Interpretations of the Gricean Maximes in the Generation of Referring Expressions. *Cognitive Science* **18**: 233-263.

- DeVault and Stone 2004. David DeVault and Matthew Stone. Interpreting vague utterances in context. In Procs. of COLING 2004, Geneva, pages 1247-1253.

- Ebeling and Gelman 1994. Ebeling, K.S., Gelman S.A. 1994. Children's use of context in interpreting "big" and "little". *Child Development* **65(4)**: 1178-1192.

- Feldman 1980. Allan M. Feldman. "Welfare Economics and Social Choice Theory". Kluwer, Boston.

- Funakoshi et al. 2004. Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, Takenobu Takunaga. Generating Referring Expressions Using Perceptual Groups. In Procs. of 3rd Int. Conf. on Natural Language Generation (INLG) 2004, Brockenhurst, UK, pages 51-60.

- Gaiffe and Romary 1997. Bertrand Gaiffe and Laurent Romary. Constraints on the Use of Language, Gesture, and Speech for Multimodal Dialogues. In Procs. of ACL workshop Referring Phenomena in a Multimedia Context and their Computational Treatment, pages 94-98.

- Goldberg et al. 1994. Eli Goldberg, Norbert Driedger, and Richard Kitteridge. Using Natural-Language Processing to Produce Weather Forecasts. IEEE *Expert* **9** (2): 45-53.

- Gorniak and Roy 2003. Peter Gorniak and Deb Roy. Understanding Complex Visually Referring Utterances. In Procs. of HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data. Edmonton, Canada, May 2003.

- Greenbaum et al. 1985. Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. A Comprehensive Grammar of the English Language. Longman, Harlow, Essex.

- Grice 1975. Paul Grice. Logic and Conversation. In P. Cole and J. Morgan (Eds.), "Syntax and Semantics: Vol 3, Speech Acts": 43-58. New York, Academic Press.

- Hermann and Deutsch 1976. T.Hermann and W.Deutsch. *Psychologie der Objektbenennung*. Huber Verlag, Bern.

- Hyde 2002. Dominic Hyde. "Sorites Paradox". The Stanford Encyclopedia of Philosophy (Fall 2002 Edition), Edward Zalta (Ed.) `http://plato.stanford.edu/archives/fall2002/entries/sorites-paradox/`.

- James et al. 1996. Glyn James, David Burley, Dick Clements, Phil Dyke, John Searl, and Jerry Wright. "Modern Engineering Mathematics". Addison-Wesley Longman Ltd., second edition. Harlow, UK.

- Kamp 1975. Hans Kamp. Two theories about adjectives. In "Semantics for natural language", ed. E. Keenan. Cambridge University Press.

- Kennedy 1999. Christopher Kennedy. *Projecting the adjective: the syntax and semantics of gradability and comparison.* PhD thesis, UC Santa Cruz.

- Krahmer and Theune 2002. Emiel Krahmer and Mariët Theune. Efficient context-sensitive generation of referring expressions. In *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, K. van Deemter and R. Kibble (eds.), CSLI Publications, CSLI, Stanford, 223–264.

- Krahmer and Van der Sluis 2003. Emiel Krahmer and Ielka van der Sluis. A New Model for Generating Multimodal Referring Expressions. In Procs. of 9th European Workshop on Natural Language Generation (ENLG-2003), Budapest, pages 47-54.

- Kyburg and Morreau (2000). Alice Kyburg and Michael Morreau. Fitting Words: vague language in context. Linguistics and Philosophy 23, pp. 577-97.

- Levelt 1989. W.J.M.Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, Mass.

- Malouf 2000. Rob Malouf. The order of prenominal adjectives in natural language generation. In procs. of ACL-2000, Hong Kong, pages 85-92.

- Masthoff 2004. Judith Masthoff. Group modeling: Selecting a sequence of television items to suit a group of viewers.User Modeling and User Adapted Interaction, 14, pp37-85.

- Mellish 2000. Chris Mellish. Understanding Shortcuts in NLG Systems. In Procs. of workshop 'Impacts in Natural Language Generation', Dagstuhl, Germany, pages 43-50.

- Nash 1950. John Nash. The Bargaining Problem. *Econometrica* **18**, 155-162.

- Peccei 1994. Jean Stilwell Peccei, "Child Language", Routledge 1994.

- Pechmann 1989. Thomas Pechmann. Incremental Speech Production and Referential Overspecification. *Linguistics* **27**: 98-110.

- Pinkal 1979. Manfred Pinkal. How to refer with vague descriptions. In R.Bäuerle, U.Egli, and A. von Stechow (Eds.) *Semantics from different points of view*. Berlin, Springer Verlag, 32-50.

- Quirk et al. 1972. Randolph Quirk, Sidney Greenbaum, and Geoffrey Leech. "A Grammar of Contemporary English". Longman, Harlow, Essex.

- Rasmusen 1989. Eric Rasmusen. *Games and Information: an Introduction to Game Theory*. Blackwell Publishing

- Reiter and Dale 2000. Ehud Reiter and Robert Dale. *Building Natural language Generation Systems.* Cambridge University Press, Cambridge, UK.

- Reiter and Sripada 2002. Ehud Reiter and Somayajulu (Yaji) Sripada. Should Corpora Texts Be Gold Standards for NLG? In Procs. of Second International Conference on Natural Language Generation (INLG-2002), pages 97-104, New York.

- Rosch 1978. Eleanor Rosch: Principles of Categorization. In: E. Rosch / B. Lloyd (eds.): Cognition and Categorization, Hillsdale, NJ: Lawrence Erlbaum, 27-48.

- Sedivy et al. 1999. Julie Sedivy, Michael Tanenhaus, Craig Chambers, and Gregory Carlson. Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109-147.

- Shaw and Hatzivassiloglou 1999. James Shaw and Vasileios Hatzivassiloglou. Ordering Among Premodifiers. In Procs. of ACL99, University of Maryland, College Park, pages 135-143.

- Sonnenschein 1982. S. Sonnenschein. The Effects of Redundant Communications on Listeners: When More is Less. *Child Development* **53**, 717-729.

- Sripada et al. 2003. Yaji Sripada, Ehud Reiter and Ian Davy. SumTime-Mousam: Configurable Marine Weather Forecast Generator. Expert Update6(3):4-10.

- Stone 2000. Matthew Stone. On Identifying Sets. In Procs. of INLG-2000, Mitzpe Ramon, pages 116-123.

- Stone and Webber 1998. Matthew Stone and Bonnie Webber. Textual Economy through Close Coupling of Syntax and Semantics. In Procs. of INLG-1998, pages 178-187.

- Thórisson 1994. Kristinn R. Thórisson. Simulated perceptual grouping: an application to human-computer interaction. Procs. of 6th Annual Conference of the Cognitive Science Society, pages 876-881.

- Toogood 1980. J.H.Toogood. What do we mean by 'usually'? *Lancet* 1980; **1**, p.1094.

- van Deemter and Odijk (1997). Kees van Deemter and Jan Odijk. Context Modeling and the Generation of Spoken Discourse. *Speech Communication* **21**, 101-121.

- van Deemter 2000. Kees van Deemter. Generating Vague Descriptions. In Procs. of Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon, pages 179-185.

- van Deemter 2002. Kees van Deemter. Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. *Computational Linguistics* **28** (1), 37-52. March 2002.

- van Deemter (2004). Kees van Deemter. Finetuning an NLG system through experiments with human subjects: the case of vague descriptions. In Procs. of 3rd Int. Conf. on Natural Language Generation (INLG-04), Brockenhurst, UK, pages 31-40.