# Computational Generation of Referring Expressions: A Survey

Emiel Krahmer*
Tilburg University

Kees van Deemter**
Aberdeen University

*This article offers a survey of computational research on referring expressions generation (REG). It introduces the REG problem and describes early work in this area, discussing what basic assumptions lie behind it, and showing how its remit has widened in recent years. We discuss computational frameworks underlying REG, and demonstrate a recent trend that seeks to link up REG algorithms with well-established Knowledge Representation traditions. Considerable attention is given to recent efforts at evaluating REG algorithms and the lessons that they allow us to learn. The article concludes with a discussion of the way forward in REG, focussing on references in larger and more realistic settings.*

## 1. Introduction

Suppose you want to point out one of the people in this scene to an addressee:

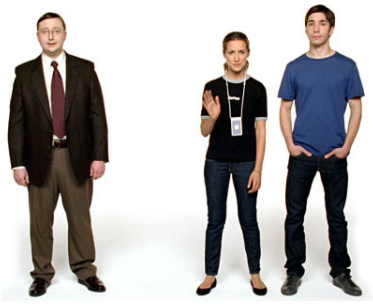


**Figure 1**
A simple visual scene.

Most speakers have no difficulty in accomplishing this task, by producing a *referring expression* such as "the man in a suit", for example. Now imagine a computer being confronted with the same task, aiming to point out individual $d_1$. Assuming it has access to a database containing all the relevant properties of the people in the scene, it needs to find some combination of properties which applies to $d_1$, and not to the other two. There is a choice though: there are many ways in which $d_1$ can be set apart from the rest ("the man on the left", "the man with the glasses", "the man with the tie"), and

* Tilburg Center for Cognition and Communication (TiCC), The Netherlands, e.j.krahmer@uvt.nl.
** Computing Science Department, University of Aberdeen, Scotland, UK, k.vdeemter@abdn.ac.uk.

the computer has to decide which of these is optimal in the given context. Moreover, optimality can mean different things. It might be thought, for instance, that references are optimal when they are minimal in length, containing just enough information to single out the target. But, as we shall see, finding minimal references is computationally expensive, and it is not necessarily what speakers do, nor what is most useful to hearers.

*So, what is Referring Expression Generation?* Referring expressions play a central role in communication, and have been studied extensively in many branches of (computational) linguistics, including Natural Language Generation (NLG). NLG is concerned with the process of automatically converting non-linguistic information (e.g., from a database) into natural language text, which is useful for practical applications ranging from generating weather forecasts to summarizing medical information (Reiter and Dale 2000). Of all the subtasks of NLG, Referring Expression Generation (REG) is among the ones that have received most scholarly attention. A survey of implemented, practical NLG systems shows that virtually all of them, regardless of their purpose, contain a REG module of some sort (Mellish et al. 2006). This is hardly surprising in view of the central role that reference plays in communication. A system providing advice about air travel (White, Clark, and Moore 2010), needs to refer to flights ("the cheapest flight", "the KLM direct flight"); a Pollen forecast system (Turner et al. 2008) needs to generate spatial descriptions for areas with low or high pollen levels ("the central belt and further North"), and a robot dialogue system that assembles construction toys together with a human user (Giuliani et al. 2010), needs to refer to the components ("insert the green bolt through the end of this red cube").

REG "is concerned with how we produce a description of an entity that enables the hearer to identify that entity in a given context" (Reiter and Dale 2000: 55). Since this can often be done in many different ways, a REG algorithm needs to make a number of choices. According to Reiter and Dale (2000), the first choice concerns what *form* of referring expression is to be used; should the target be referred to, for instance, using its proper name, a pronoun ("he") or a description ("the man with the tie"). Proper names have limited applicability because many domain objects do not have a name that is in common usage. For pronoun generation, a simple but conservative rule is discussed by Reiter and Dale (2000), similar to one proposed by Dale (1989): use a pronoun if the target was mentioned in the previous sentence, and if this sentence contained no reference to any other entity of the same gender (p. 150-151). Reiter and Dale (2000) concentrate mostly on the generation of descriptions. If the NLG system decides to generate a description, two choices need to be made: which set of properties distinguishes the target (content selection), and how can the selected properties be turned into natural language (linguistic realisation). Content selection is a complex balancing act: we need to say enough to enable identification of the intended referent, but not too much. A selection of information needs to be made, and this needs to be done quickly. Reiter and Dale discuss various strategies that try to manage this balancing act, based on Dale and Reiter (1995), an early survey article that summarises and compares various influential algorithms for the generation of descriptions.

*Why a survey on REG, and how to read it?* REG, like NLG in general, has changed considerably from the overviews presented in Dale and Reiter (1995) and Reiter and Dale (2000), owing largely to an increased use of empirical data, and a widening of the class of referring expressions studied. Moreover, a gradual shift has taken place towards extended application domains, different input and output formats, and more flexible interactions with the user, and this shift is starting to necessitate the use of new

REG techniques. Examples include recent systems in areas such as weather forecasting (Turner, Sripada, and Reiter 2009) and medical care (Portet et al. 2009) where complex references to spatial regions and time periods abound. The results of recent REG research are scattered over proceedings, books and journals. The current survey offers a compact overview of the progress in this area and an assessment of the state of the art.

The concept of reference is difficult to pin down (Searle 1969; Abbott 2010). Searle suggests that the proper approach is "to examine those cases which constitute the center of variation of the concept of referring and then examine the borderline cases in light of similarities and differences from the paradigms" (Searle 1969: 26-27). This is precisely what we shall do in this survey. The "paradigms" of reference in Reiter and Dale (2000) are *definite descriptions* whose primary purpose it is to *identify their referent*, and to do this *without reliance on the linguistic context* of the expression. Most recent REG research subscribes to this view as well. We shall often have occasion to discuss other types of expressions, but our main focus will be on these paradigmatic cases. To do full justice to indefinite or attributive descriptions, proper names, and personal pronouns would, in our view, require a separate survey.

In Section 2 we offer a brief overview of REG research up to 2000, discussing some classic algorithms. Next, we zoom in on the new directions in which recent work has taken REG research: extension of the coverage of algorithms, to include, for example, vague, relational and plural descriptions (Section 3), exploration of different computational frameworks, such as Graph Theory and Description Logic (Section 4) and collection of data and evaluation of REG algorithms (Section 5). Section 6 highlights open questions and avenues for future work. Section 7 summarises our findings.

## 2. A very short history of pre-2000 REG research

The current survey focusses primarily on the progress in REG research in the 21st century, but it is important to have a basic insight into pre-2000 REG research and how it laid the foundation for much of the current work.

### 2.1 First beginnings

REG can be traced back to the earliest days of Natural Language Processing; Winograd (1972) (Section 8.3.3, Naming Objects and Events), for example, sketches a primitive "incremental" REG algorithm, used in his SHRDLU program. In the 1980s, researchers such as Appelt and Kronfeld set themselves the ambitious task of modelling the human capacity for producing and understanding referring expressions in programs such as KAMP and BERTRAND (Appelt 1985; Appelt and Kronfeld 1987; Kronfeld 1990). They argued that referring expressions should be studied as part of a larger speech act. KAMP (Appelt 1985), for example, was conceived as a general utterance planning system, building on Cohen and Levesque's (1985) formal speech act theory. It used logical axioms and a theorem prover to simulate an agent planning instructions such as "use the wheelpuller to remove the flywheel", which contains two referring expressions, as part of a larger utterance.

Like many of their contemporaries, Appelt and Kronfeld's hoped to gain insight into the complexities of human communication. Doug Appelt (p.c.): "(...) the research themes that originally motivated our work on generation were the outgrowth of the methodology in both linguistics and computational linguistics at the time that research progress was best made by investigating hard, anomalous cases that pose difficulties for conventional accounts." Their broad focus allowed these researchers to recognise

that although referring expressions may have identification of the referent as their main goal, a referring expression can also *add* information about a target. By pointing to a tool on a table, while saying "the wheelpuller", the descriptive content of the referring expression may serve to inform the hearer about the function of the tool (Appelt and Kronfeld 1987). They also observed that referring expressions need to be sensitive to the communicative context in which they are used and that they should be consistent with the Gricean maxims (see below), which militate against overly elaborate referring expressions (Appelt 1985)

It is remarkably difficult, after 20 years, to find out how these programs actually worked, since code was lost and much of what was written about them is pitched at a high level of abstraction. Appelt and Kronfeld were primarily interested in difficult questions about human communication, but they were sometimes tantalisingly brief about humbler matters. Here, for instance, is how Appelt (1985) (p. 21) explains how KAMP would attempt to identify a referent:

> "KAMP chooses a set of basic descriptors when planning a describe action to minimise both the number of descriptors chosen, and the amount of effort required to plan the description. Choosing a provably minimal description requires an inordinate amount of effort and contributes nothing to the success of the action. KAMP chooses a set of descriptors by first choosing a *basic category* descriptor (see [Rosch 1978]) for the intended concept, and then adding descriptors from those facts about the object that are mutually known by the speaker and the hearer, subject to the constraint that they are all linguistically realizable in the current noun phrase, until the concept has been uniquely identified. (. . . ) Some psychological evidence suggests the validity of the minimal description strategy; however, one does not have to examine very many dialogues to find counter-examples to the hypothesis that people always produce minimal descriptions."

This quote contains the seeds of much later work in REG, given its skepticism about the naturalness of minimal descriptions, its use of Rosch (1978)-style basic categories, and its acknowledgment of the role of computational complexity. Broadly speaking, it suggests an *incremental* generation strategy, compatible with the ones described below, although it is uncertain what exactly was implemented. In recent years, the Appelt-Kronfeld line of research has largely given way to a new research tradition which focussed away from the full complexity of human communication, with notable exceptions such as Heeman and Hirst (1995), Stone and Webber (1998), O'Donnell, Cheng and Hitzeman (1998), and Koller and Stone (2007).

### 2.2 Generating distinguishing descriptions

In the early nineties a new approach to REG started gaining currency, when Dale and Reiter re-focussed on the problem of determining what properties a referring expression should use if identification of the referent is the central goal (Dale 1992, 1989; Reiter 1990; Reiter and Dale 1992). This line of work culminated in the seminal Dale and Reiter (1995). Like Appelt (1985), Dale and Reiter are concerned with the link between the Gricean maxims and the generation of referring expressions. They discuss the following pair of examples:

(1) Sit by *the table*.
(2) Sit by *the brown wooden table*.

**Table 1**
Tabular representation of some information in our example scene.

| Object | type | clothing | position |
|---|---|---|---|
| $d_1$ | man | wearing suit | left |
| $d_2$ | woman | wearing t-shirt | middle |
| $d_3$ | man | wearing t-shirt | right |

In a situation where there is only one table, which happens to be brown and wooden, both the descriptions in (1) and (2) would successfully refer to their target. However, if you hear (2) you might make the additional inference that it is significant to know that the table is brown and wooden; why else would the speaker mention these properties? If the speaker merely wanted to refer to the table, your inference would be an (incorrect) "conversational implicature", caused by the speaker's violation of Grice's (1975: 45) Maxim of Quantity ("Do not make your contribution more informative than is required."). Dale and Reiter (1995) ask how we can efficiently compute which properties to include in a description, such that it successfully identifies the target while not triggering false conversational implicatures. For this, they zoom in on a relatively straightforward problem definition, and compare a number of concise, well-defined algorithms solving the problem.

*Problem definition.* Dale and Reiter (1995) formulate the REG problem as follows. Assume we have a finite domain $D$ of objects with attributes $A$. In our example scene (Figure 1), $D = \{d_1, d_2, d_3\}$ and $A = \{\textsf{type}, \textsf{clothing}, \textsf{position}, \ldots\}$. The type attribute has a special status in Dale and Reiter (1995) since it represents the semantic content of the head noun. Alternatively, we could have defined an attribute gender, stating that it should be realised as the head noun of a description. Typically, domains are represented in a knowledge base such as Table 1, where different values are clustered together because they are associated with the same attribute. Left, right and middle, for example, belong to the attribute position, and are said to be three *values* that this attribute can take. The objects of which a given attribute–value combination (or "property") is true are said to form its *denotation*. Sometimes we will drop the attribute, writing man, rather than $\langle \textsf{type}, \textsf{man} \rangle$, for instance.

The REG task is now defined by Dale and Reiter (1995) through what may be called *identification* of the target: given a **target** (or referent) object $r \in D$, find a set of attribute–value pairs $L$ whose conjunction is true of the target but not of any of the **distractors** (i.e., $D - \{r\}$, the domain objects different from the target). $L$ is called a *distinguishing description* of the target. In our simple example, suppose that $\{d_1\}$ is the target (and hence $\{d_2, d_3\}$ the set of distractors), then $L$ could, for example, be either $\{\langle \textsf{type}, \textsf{man} \rangle, \langle \textsf{clothing}, \textsf{wearing suit} \rangle\}$ or $\{\langle \textsf{type}, \textsf{man} \rangle, \langle \textsf{position}, \textsf{left} \rangle\}$, which could be realised as "the man wearing a suit" or "the man to the left". If identification were all that counted, a simple, fast, and fault-proof REG strategy would be to *conjoin all the properties of the referent*: this conjunction will identify the referent if it can be identified at all. In practice, Dale and Reiter, and others in their wake, include an additional, constraint which is often left implicit: that the referring expressions generated *should be as similar to human-produced ones* as possible. In the Evaluation and Conclusion sections, we return to this "human-likeness" constraint (and to variations on the same theme).

*Full Brevity and Greedy Heuristic.* Dale and Reiter (1995) discuss various algorithms which solve the REG task. One of these is the **Full Brevity** algorithm (Dale 1989) which deals with the problem of avoiding false conversational implicatures in a radical way, by always generating the shortest possible distinguishing description. Originally, the Full Brevity algorithm was meant to generate both initial and subsequent referring expressions, by relying on a previous step that determines the distractor set based on which objects are currently salient. Given this set, it first checks whether there is a single property of the target that rules out all distractors. If this fails, it considers all possible combinations of *two* properties, and so on:

1. Look for a description $L$ that distinguishes target $r$ using *one* property. If success then return $L$. Else go to 2.
2. Look for a description $L$ that distinguishes target $r$ using *two* properties. If success then return $L$. Else go to 3.
3. *Etcetera*

Unfortunately, there are two problems with this approach. First, the problem of finding a shortest distinguishing description has a high complexity (it is *NP hard*, see e.g., Garey and Johnson (1979)) and hence is computationally very expensive, making it prohibitively slow for large domains and descriptions. Second, Dale and Reiter note that human speakers routinely produce descriptions that are *not* minimal. This is confirmed by a substantial body of psycholinguistic research (Olson 1970; Sonnenschein 1984; Pechmann 1989; Engelhardt, Bailey, and Ferreira 2006).

An approximation of Full Brevity is the **Greedy Heuristic** algorithm (Dale 1989, 1992), which iteratively selects the property which rules out most of the distractors not previously ruled out, incrementally augmenting the description based on what property has most discriminatory power at each stage (as a result, it does not always generate descriptions of minimal size). The Greedy Heuristic algorithm is a more efficient algorithm than the Full Brevity one, but it was soon eclipsed by another algorithm (Reiter and Dale 1992; Dale and Reiter 1995), which turned out to be the most influential algorithm of the pre-2000 era. It is this later algorithm that came to be known as "the" **Incremental Algorithm** (IA).

*The Incremental Algorithm.* The basic idea underlying the IA is that speakers "prefer" certain properties over others when referring to objects, an intuition supported by the experimental work of, for instance, Pechmann (1989). Suppose you want to refer to a person 10 metres away from you. You might mention the person's gender. If this is insufficient to single out the referent, you might be more likely to make use of the colour of the person's coat than to the colour of her eyes. Less preferred attributes, such as eye colour, are only considered if other attributes do not suffice. It is this intuition of a *preference* order between attributes that the IA exploits. By making this order a parameter of the algorithm, a distinction can be made between domain/genre dependent knowledge (the preferences), and a domain-independent search strategy.

As in the Greedy Heuristic algorithm, descriptions are constructed incrementally; but unlike the Greedy Heuristic, the IA checks attributes in a fixed order. By grouping properties into attributes, Dale and Reiter predict that all values of a given attribute have the same preference order. Ordering attributes rather than values, may be disadvantageous, however. A simple shape (e.g., a circle), or a size that is unusual for its target (e.g., a tiny whale) might be preferred over a subtle colour (purplish grey). Also,

```
1.  IncrementalAlgorithm ({r}, D, Pref){
2.  L ← ∅
3.  C ← D − {r}
4.  for each A_i in list Pref do
5.      V = Value(r, A_i)
6.      if C ∩ RulesOut(⟨A_i, V⟩) ≠ ∅
7.      then L ← L ∪ {⟨A_i, V⟩}
8.           C ← C − RulesOut(⟨A_i, V⟩)
9.      endif
10.     if C = ∅
11.     then return L
12.     endif
13. return failure }
```

**Figure 2**
Sketch of the core Incremental Algorithm

some values of a given attribute might be difficult to express, and "dispreferred" for this reason (kind of like a ufo shape with a christmas tree sticking out the side)

Figure 2 contains a sketch of the IA in pseudo code. It takes as input a target object $r$, a domain $D$, consisting of a collection of domain objects, and a domain-specific list of preferred attributes *Pref* (1). Suppose we apply the IA to $d_1$ of our example scene, assuming that *Pref* = type > clothing > position. The description $L$ is initialised as the empty set (2), and the context set $C$ of distractors (from which $d_1$ needs to be distinguished) is initialised as $D - \{d_1\}$ (3). The algorithm then iterates through the list of preferred attributes (4), for each one looking up the target's value (5), and checking whether this attribute–value pair rules out any of the distractors not ruled out so far (6). The function RulesOut $(\langle A_i, V\rangle)$ returns the set of objects which have a different value for attribute $A_i$ than the target object has. If one or more distractors are ruled out, the attribute–value pair $\langle A_i, V\rangle$ is added to the description under construction (7) and a new set of distractors is computed (8). The first attribute to be considered is type, for which $d_1$ has the value man. This would rule out $d_2$, the only woman in our domain, and hence the attribute–value pair $\langle$type, man$\rangle$ is added to $L$. The new set of distractors is $C = \{d_3\}$, and the next attribute (clothing) is tried. Our target is wearing suit, and the remaining distractor is not, so the attribute–value pair $\langle$clothing, wearing suit$\rangle$ is included as well. At this point all distractors are ruled out (10), a set of properties has been found which uniquely characterise the target $\{\langle$type, man$\rangle, \langle$clothing, wearing suit$\rangle\}$ ("the man wearing a suit"), and we are done (11). If we had reached the end of the list without ruling out all distractors, the algorithm would have failed (13): no distinguishing description for our target was found.

The sketch in Figure 2 simplifies the original algorithm in a number of respects. First, Dale and Reiter always include the type attribute, even if it does not rule out any distractors, because speakers use type information in virtually all their descriptions. Second, the original algorithm checks, via a function called UserKnows, whether a given property is in the common ground, to prevent the selection of properties which the addressee might not understand. Unlike Appelt and Kronfeld, who discuss detailed examples that hinge on differences in common ground, Dale and Reiter (1995) treat UserKnows as a function that returns "true" for each true proposition, assuming that all relevant information is shared. Third, the IA can take some ontological information

into account via subsumption hierarchies. For instance, in a dog-and-cat domain, a pet may be of the chihuahua type, but chihuahua is subsumed by dog, and dog in turn is subsumed by animal. A special value in such a subsumption hierarchy is reserved for the so-called basic level values (Rosch 1978); dog in this example. If an attribute comes with a subsumption hierarchy, the IA first computes the best value for that attribute, which is defined as the value closest to the basic level value, such that there is no *more specific* value that rules out *more* distractors. In other words, the IA favours dog over chihuahua, unless the latter rules out more distractors.

The IA is conceptually straightforward and easy to implement. In addition, it is computationally efficient, with *polynomial* complexity: its worst-case run time is a constant function of the total number of attribute–value combinations available. This computational efficiency is due to the fact that the algorithm does not perform backtracking: once a property has been selected, it is included in the final referring expression, even if later additions render it superfluous. As a result, the final description may contain redundant properties. Far from seeing this as a weakness, Dale & Reiter (1995: 19) point out that this makes the IA less "psycholinguistically implausible" than its competitors. It is interesting to observe that while Dale and Reiter (1995) discuss the theoretical complexity of the various algorithms in detail, later research has tended to attach more importance to empirical evaluation of the generated expressions (Section 5).

### 2.3 Discussion

Appelt and Kronfeld's work, founded on the assumption that REG should be seen as part of a comprehensive model of communication, started to lose some of its appeal in the early nineties, because it was at odds with the emerging research ethos in computational linguistics that stressed simple, well-defined problems allowing for measurable results. The way current REG systems are shaped is largely due to developments summarised in Dale and Reiter (1995), which focusses on a specific aspect of REG, namely determining which properties serve to identify some target referent. Dale and Reiter's work aimed for generating human-like descriptions, but was not coupled with systematic investigation of data.

*REG as search.* The algorithms discussed by Dale and Reiter (1995) can be seen as different instantiations of a general search algorithm (Bohnet and Dale 2005; Gatt 2007). They all basically search through the same space of states, each consisting of three components: a description that is true of the target, a set of distractors, and a set of properties of the target that have not yet been considered. The initial state can be formalised as the triple $\langle \emptyset, C, P \rangle$ (no description for the target has been constructed, no distractors have been ruled out, and all properties $P$ of the target are still available), and the goal state as $\langle L, \emptyset, P' \rangle$, for certain $L$ and $P'$: a description $L$ has been found, which is distinguishing – the set of distractors is empty. All other states in the search space are intermediate ones, through which an algorithm might move depending on its search strategy. For instance, when searching for a distinguishing description for $d_1$ in our example domain, an intermediate state could be $s = \langle \{\langle \mathsf{type}, \mathsf{man} \rangle\}, \{d_3\}, \{\langle \mathsf{clothing}, \mathsf{wearing\ suit} \rangle, \langle \mathsf{position}, \mathsf{left} \rangle\} \rangle$.

The algorithms discussed earlier differ in terms of their so-called *expand* method, determining how new states are created, and their *queue* method, which determines the order in which these states are visited (i.e., how states are inserted into the queue). Full Brevity, for example, uses an expand-method that creates a new state for each attribute of the target not checked before (as long as it rules out at least one distractor).

Starting from the initial state and applied to our example domain, this expand-method would result in 3 new states, creating descriptions including type, clothing and position information respectively. These states would be checked using a queue method which is breadth-first. The IA, by contrast, uses a different expand-method, each time creating a single new state in accordance with the pre-determined preference order. Thus, in the initial state, and assuming (as before) that type is the most preferred attribute, the expand method would create a single new state: $s$ above. Since there is always only one new state, the queue method is trivial.

*Limitations of pre-2000 REG.* In the IA and related algorithms, the focus is on efficiently computing which properties to use in a distinguishing description. However, there are a number of implicit simplifications in the way the task is framed. (1) The target is always just one object, not a larger set (hence, plural noun phrases are not generated). (2) The algorithms all assume a very simple kind of knowledge representation, consisting of a set of *atomic* propositions. Negated propositions are only represented indirectly, via the *Closed World Assumption*, so an atomic proposition that is not explicitly listed in the database is false. (3) Properties are always "crisp", never vague. Vague properties such as small and large are treated as Boolean properties, which do not allow borderline cases and which keep the same denotation, regardless of the context in which they are used. (4) All objects in the domain are assumed to be equally salient, which implies that *all* distractors have to be removed, even those having a very low salience. (5) The full REG task includes first determining which properties to include in a description, *and* then providing a surface realisation in natural language of the selected properties. The second stage is not discussed, nor is the relation with the first. As we shall, a substantial part of recent REG research is dedicated to lifting one or more of these simplifying assumptions. Other limitations are still firmly in place (as we shall discuss in Section 6).

## 3. Extending the coverage

### 3.1 Reference to sets

Until recently, REG algorithms aimed to produce references to a single object. But references to sets are ubiquitous in most text genres. In simple cases, it takes only a slight modification to allow classic REG algorithms to refer to sets. The IA, for example, can be seen as referring to the singleton *set* $\{r\}$ that contains the target $r$ and nothing else. If in line 1 (Figure 2), $\{r\}$ is replaced by an arbitrary set $S$, and line 3 is modified as saying $C \leftarrow D - S$ instead of $C \leftarrow D - \{r\}$, then the algorithm produces a description that applies to all elements of $S$. Thus, it is easy enough to let these algorithms produce expressions like "the men" or "the t-shirt wearers", to identify $\{d_1, d_3\}$ and $\{d_2, d_3\}$ respectively. Unfortunately, things are not always so simple. What if we need to refer to the set $\{d_1, d_2\}$? Based on the properties in Table 1 alone this is not possible, because $d_1$ and $d_2$ have no properties in common. The natural solution is to treat the target set as the union of two smaller sets, $\{d_1\} \cup \{d_2\}$, and refer to both sets separately (e.g., "the man who wears a suit, and the woman"). Once unions are used, it becomes natural to allow set complementation as well, as in "the people who are *not* on the right". Note that set complementation may also be useful for single referents. Consider a situation where all cats except one are owned by Mary, while the owner of the remaining one is unknown or non-existent. Complementation allows one to refer to "the cat *not* owned by Mary". Henceforth we shall refer to the resulting descriptions as *Boolean*.

1. **[Length 1.]** Run IA using all properties of the form $P_{+/-}$
   If success then return $L$ else goto (2) to add new properties to $L$.
2. **[Length 2.]** Run IA using all properties of the form $P_{+/-} \cup P_{+/-}$
   If success then return $L$ else goto (3) to add new properties to $L$.
3. **[Length 3.]** Run IA using all properties of the form $P_{+/-} \cup P_{+/-} \cup P_{+/-}$
   If success then return $L$ else goto (4) to add new properties to $L$.
4. *Etcetera*, up to unions of length $n$.

**Figure 3**
Outline of the first stage of van Deemter's (2002) Boolean REG algorithm.

As part of a more general logical analysis of the IA, van Deemter (2002) made a first stab at producing Boolean descriptions, using a two-stage algorithm whose first stage is a generalisation of the IA, and whose second stage involves the optimisation of the possibly lengthy expressions produced by the first phase. The resulting algorithm is *logically complete* in the following sense: if a given set can be described at all using the properties available then this algorithm will find such a description.

The first stage of the algorithm starts by conjoining properties (man, left) in the familiar manner of the IA; if this does not suffice for singling out the target set then the same incremental process continues with unions of two properties (e.g., man $\cup$ woman, middle $\cup$ left), then with unions of three properties (e.g., man $\cup$ wearing suit $\cup$ woman), and so on. The algorithm terminates when the referent (individual or set) is identified (success) or when all combinations of properties have been considered (failure). Figure 3 depicts this in schematic form, where $n$ represents the total number of properties in the domain, and $P_{+/-}$ denotes the set of all *literals* (atomic properties such as man, and their complements $\neg$man). Step (1) generalises the original IA allowing for negated properties and target sets. As before, $L$ is the description under construction. It will consist of intersections of unions of literals such as (woman $\cup$ man) $\cap$ (woman $\cup \neg$wearing suit) (in other words, $L$ is in Conjunctive Normal Form, CNF).

Note that this first stage is not only incremental at each of its $n$ steps, but also as a whole: once a property has been added to the description, later steps will not remove it. This can lead to redundancies, even more than in the original IA. The second stage removes the most blatant of these, but only where the redundancy exists as a matter of logic, rather than world knowledge. Suppose, for example, that step 2 selects the properties $P \cup S$ and $P \cup R$, ruling out all distractors. $L$ now takes the form $(P \cup S) \cap (P \cup R)$ (e.g., "the things that are both (women or men) and (women or wearing suits)"). The second phase uses logic optimisation techniques (originally designed for the minimisation of digital circuits (McCluskey 1965)) to simplify this to $P \cup (S \cap R)$ ("the women, and the men wearing suits").

*Variations and extensions.* Gardent (2002) drew attention to situations where this proposal produces unacceptably lengthy descriptions; suppose, for example, the algorithm accumulates numerous properties during steps 1 and 2, before finding one complex property (a union of three properties) during step 3 which, on its own would have sufficed to identify the referent. This will make the description generated much lengthier than necessary, because the properties from steps 1 and 2 are now superfluous. Gardent's take on this problem amounts to a reinstatement of Full Brevity embedded

in a reformulation of REG as a constraint satisfaction problem (see Section 4.2). The existence of fast implementations for constraint satisfaction alleviates the problems with computational tractability to a considerable extent. But by re-instating Full Brevity, algorithms like Gardent's could run into the empirical problems noted by Dale and Reiter, given that human speakers frequently produce non-minimal descriptions (see Gatt (2007) for evidence pertaining to plurals).

Horacek (2004) makes a case for descriptions in Disjunctive Normal Form (DNF; unions of intersections of literals). Horacek's algorithm first generates descriptions in CNF, then convert these into DNF, skipping superfluous disjuncts. Consider our example domain (Table 1). To refer to $\{d_1, d_2\}$, a CNF-oriented algorithm might generate $(\mathsf{man} \cup \mathsf{woman}) \cap (\mathsf{left} \cup \mathsf{middle})$ ("the people who are on the left or middle"). Horacek converts this, first, into DNF: $(\mathsf{man} \cap \mathsf{left}) \cup (\mathsf{woman} \cap \mathsf{middle}) \cup (\mathsf{man} \cap \mathsf{middle}) \cup (\mathsf{woman} \cap \mathsf{left})$, after which the last two disjuncts are dropped, because there are no men in the middle, and no women on the left. The outcome could be worded as "the man on the left and the woman in the middle". Later work has tended to agree with Horacek in opting for DNF instead of CNF (Gatt 2007; Khan, van Deemter, and Ritchie 2008).

*Perspective and coherence.* Recent studies have started to bring data-oriented methods to the generation of references to sets (Gatt 2007; Gatt and van Deemter 2007; Khan, van Deemter, and Ritchie 2008). One finding is that referring expressions benefit from a "coherent" perspective. For example, "the Italian and the Greek" is normally a better way to refer to two people than "the Italian and the cook", since the former is generated from one coherent perspective (i.e., nationalities). Two questions need to be addressed, however. First, how should coherence be defined? Gatt (2007) opted for a definition that assesses the coherence of a combination of properties using corpus-based frequencies as defined by Kilgarriff (2003) (which is based on Lin (1998)). This choice was supported by a range of experiments (although the success of the approach is less well attested for referring expressions that contain adjectives). Secondly, what if full coherence can only be achieved at the expense of brevity? Suppose a domain contains one Italian and two Greeks. One of the Greeks is a cook, while the other Greek and the Italian are both IT consultants. If this is all that is known, the generator faces a choice between either generating a description that is fully coherent but unnecessarily lengthy ("the Italian IT consultant and the Greek cook"), or brief but incoherent ("The Italian and the cook"). Simply saying "The Italian and the Greek" would not be distinguishing. In such cases, coherence becomes a tricky, and computationally complex, optimisation problem (Gatt 2007; Gatt and van Deemter 2007).

*Collective plurals.* Reference to sets is a rich topic, where many issues on the borderline between theoretical, computational, and experimental linguistics are waiting to be explored. Most computational proposals, so far, use properties that apply to individual objects. To refer to a set, in this view, is to say things that are true of each member of the set. Such references may be contrasted with *collective* ones (e.g., "the lines that run parallel to each other", "the group of 4 people") whose semantics is known to throw up many problems (see e.g., Scha and Stallard (1988) or Lønning (1997)). For initial ideas about the generation of collective plurals, we refer to Stone (2000).

### 3.2 Relational descriptions

Another important limitation of most early REG algorithms is that they are restricted to one-place predicates (e.g., "being a man"), instead of relations involving two or more arguments. Even a property like "wearing a suit" is modelled as if it were simply a one-place predicate without internal structure (instead of a relation between a person and a piece of clothing). This means that the algorithms in question are unable to identify one object via another, as when we say "the man who wears a suit that was bought by a woman who lives above the supermarket", and so on.

One early paper does discuss *relational* descriptions, making a number of important observations about them (Dale and Haddock 1991). First, it is possible to identify an object through its relations to other objects without identifying each of these objects separately. Consider a situation involving two cups and two tables, where one cup is on one of the tables. In this situation, neither "the cup" nor "the table" is distinguishing, but "the cup on the table" succeeds in identifying one of the two cups. Secondly, descriptions of this kind can have any level of 'depth': in a complex situation, one might say "the white cup on the red table in the kitchen", and so on. To be avoided, however, are the kinds of repetitions that can arise from descriptive *loops*, since these do not add information. It would, for example, be useless to describe a cup as "the cup to the left of the saucer to the right of the cup to the left of the saucer ...". We shall return to this issue in Section 4, where we shall ask how suitable each of a number of representational frameworks is for the proper treatment of relational descriptions.

Various researchers have attempted to extend the IA by allowing relational descriptions (Horacek 1996; Krahmer and Theune 2002; Kelleher and Kruijff 2006), often based on the assumption that relational properties (like "$x$ is on $y$") are less preferred than non-relational ones (like "$x$ is white"). If a relation is required to distinguish the target $x$, the basic algorithm is applied iteratively to $y$. It seems, however, that these attempts were only partly successful. One of the basic problems is that relational descriptions – just like references to sets, but for different reasons – do not seem to fit in well with an incremental generation strategy. In addition, it is far from clear that relational properties are always less preferred than non-relational ones (Viethen and Dale 2008). Viethen and Dale suggest that even in simple scenes, where objects can easily be distinguished without relations, participants still use relations frequently (in about one third of the trials). We return to this in Section 5.

On balance, it appears that the place of relations in reference is only partly understood, with much of the iceberg still under water. If 2-place relations can play a role in REG, then surely so can $n$-place relations for larger $n$, as when we say "the city that lies in between the mountains and the sea" ($n = 3$). No existing proposal has addressed $n$-place relations in general, however. Moreover, human speakers can identify a man as the man who "kissed *all* women", "*only* women", or "*two* women". The proposals discussed so far do not cover such quantified relations (but see Ren, van Deemter, and Pan (2010)).

### 3.3 Context-dependency, vagueness and gradability

So far we assumed that properties have a crisply defined meaning, which is fixed, regardless of the context in which they are used. But many properties fail to fit this mould. Consider the properties young and old, for example. In Figure 1, it is the leftmost male who looks the older of the two. But if we add an old-age pensioner to the scene then suddenly *he* is the most obvious target of expressions like "the older man" or

"the old man". Whether someone counts as old or not, in other words, depends on what other people he is compared with: being old is a context-dependent property. The concept of being "on the left" is context-dependent too: suppose we add five people to the right of the young man in Figure 1; now all three characters originally depicted are suddenly on the left, including the man in the t-shirt who started out on the right.

Concepts like "old" and "left" involve comparisons between objects. Therefore, if the knowledge base changes, the objects' descriptions may change as well. But even if the knowledge base is kept constant, the referent may have to be compared against different objects, depending on the words in the expression. The word "short" in "John is a short basketball player", for example, compares John's height with that of the other basketball players, whereas "John is a short man" compares its referent with all the other men, resulting in different standards for what it means to be short.

"Old" and "short' are not only context dependent but also *gradable*, meaning that you can be more or less of it (older, younger, shorter, taller) (Quirk et al. 1980). Gradable words are extremely frequent, and in many NLG systems they are of great importance, particularly in those that have numerical input, for example in weather forecasting (Goldberg, Driedger, and Kittredge 1994) or medical decision support (Portet et al. 2009). In addition to being context dependent, they are also *vague*, in the sense that they allow borderline cases. Some people may be clearly young, others clearly not, but there are borderline cases in between for whom it is not quite clear whether they were included. Context can help to diminish the problem, but it won't go away: in the expression "short basketball player", the noun gives additional information about the intended age range, but borderline cases still exist.

*Generating vague references.* REG, as we know it, lets generation start from a Knowledge Base (KB) whose facts do not change as a function of context. This means that context-dependent properties like a person's height need to be stored in the KB in a manner that does not depend on other facts. It is possible to deal with size adjectives in a principled way, by letting one's KB contain a height attribute with numerical values. Our running example can be augmented by giving each of the three people a precise height, for example: $\mathsf{height}(d_1) = 170cm$, $\mathsf{height}(d_2) = 180cm$ and $\mathsf{height}(d_3) = 180cm$ (here the height of the woman $d_2$ has been increased for illustrative purposes). Now imagine we want to refer to $d_3$. This target can be distinguished by the set of two properties $\{\mathsf{man}, \mathsf{height} = 180cm\}$. Descriptions of this kind can be produced by means of any of the classic REG algorithms.

Given that type and height identify the referent uniquely, this set of properties can be realised simply as "the man who is 180cm tall". But other possibilities exist. Given that $180cm$ is the greatest height of all men in this KB, the set of properties can be converted into $\{\mathsf{man}, \mathsf{height} = \mathsf{maximum}\}$, where the exact height has been pruned away. The new description can be realised as "the tallest man" or simply as "the tall man" (provided the referent's height exceeds a certain minimum value). The algorithm becomes more complicated when *sets* are referred to (because the elements of the target set may not all have the same heights), or when two or more gradable properties are combined (as in "the strong, tall man in the expensive car") (van Deemter 2006).

*Variations and extensions.* Horacek (2005) integrates vagueness with other types of uncertainty. Horacek could be said to depict a REG algorithm as a gambler who wants to maximise the chance of the referent being identified on the basis of the generated expression. Other things being equal, for example, it may be safer to identify a dog as being "owned by John", than as being "tall", because the latter involves

borderline cases. A similar approach can be applied to perceptual uncertainty (as when it is uncertain whether the hearer will be able to observe a certain property), or to the uncertainty associated with little-known words (e.g., will the hearer know what a basset hound is?) Quantifying all types of uncertainties could prove problematic in practice, yet by portraying a generator as, essentially, a gambler, Horacek has highlighted an important aspect of reference generation which had so far been ignored. Crucially, his approach makes the success of a description a matter of degrees.

The idea that referential success is a matter of degrees appears to be confirmed by recent applications of REG to geo-spatial data. Here there tend to arise situations in which it is simply not feasible to produce a referring expression that identifies its target with absolute precision (though good approximations may exist). Once again, the degree of success of a referring expression becomes gradable. Suppose you were asked to describe that area of Scotland where the temperature is expected to fall below zero on a given night, based on some computer forecast of the weather. Even if we assume that this is a well-defined area with crisp boundaries, it is not feasible to identify the area precisely, because listing all the thousands of data points that make up the area separately is hardly an option. Various approximations are possible, including:

(3) *Roads above 500 metres* will be icy.
(4) *Roads in the Highlands* will be icy.

Descriptions of this kind are generated by a system for road gritting, where the decision which roads to treat with salt depends on the description generated by the system (Turner, Sripada, and Reiter 2009): roads where temperatures are predicted to be icy should be treated with salt; others should not. The two descriptions above are arguably only partially successful in singling out the target area. Generally speaking, one can distinguish between false positives and false negatives: the former are roads that are covered by the description but should not be (because the temperature there is not predicted to fall below zero), the latter are icy roads that will be left un-gritted. Turner and colleagues decided that it would be unacceptable to have even one false negative. In other situations, safety (from accidents) and environmental damage (through salt) might be traded off in different ways, for example by associating a finite *cost* with each false positive and a possibly different cost with each false negative, and choosing the description that is associated with the lowest total cost (van Deemter 2010: 253–254). Again, a crucial and difficult part is to come up with the right cost figures.

### 3.4 Degrees of salience and the generation of pronouns

When we speak about the world around us, we do not pay equal attention to all the objects in it. In a novel, for example, a sentence like "Smiley saw the man approaching" does not mean that Smiley saw the only man: it simply means that Smiley saw the man who is most salient at this stage of the novel. Passonneau (1996) and Jordan (2000) have shown how algorithms such as the IA may produce reasonable referring expressions "in context", by limiting the set of salient objects in some sensible way, for example, to those objects mentioned in the previous utterance. Salience, in these works, was treated as a two-valued, "black-or-white" concept. But perhaps it is more natural to think of salience – just like height or age – as coming in degrees. Existing theories of linguistic salience do not merely separate what is salient from what is not. They assign referents to different salience bands, based on factors such as recency of mention and

syntactic structure (Gundel, Hedberg, and Zacharski 1993; Hajičová 1993; Grosz, Joshi, and Weinstein 1995).

*Salience and context-sensitive REG.* Early REG algorithms (Kronfeld 1990; Dale and Reiter 1995) assumed that salience could be modelled by means of a *focus stack* (Grosz and Sidner 1986): a referring expression is taken to refer to the highest element on the stack that matches its description (see also DeVault (2004)). Krahmer and Theune (2002) argue that the focus stack approach is not flexible enough for context-sensitive generation of descriptions. They propose to assign individual *salience weights* (*sw*s) to the objects in the domain, and to reinterpret referring expressions like *the man* as referring to the currently most salient man. Once such a gradable notion of salience is adopted, we are back in the territory of Section 3.3. One simple way to generate context-sensitive referring expressions is to keep the algorithm of Figure 2 exactly as it is, but to limit the set of distractors to only those domain elements whose salience weight is at least as high as that of the target $r$. Line 3 (Figure 2) becomes:

3′. $C \leftarrow \{x \mid sw(x) \geq sw(r)\} - \{r\}$

To see how this works, consider the knowledge base of Table 1 once again, assuming that $sw(d_1) = sw(d_2) = 10$, while $sw(d_3) = 0$ ($d_1$ and $d_2$ are salient, for example, because they were just talked about, and $d_3$ was not). Suppose we keep the same domain and preference order as before. Now if $d_1$ is the target, then, according to the new definition 3′, $C = \{d_1, d_2\} - \{d_1\} = \{d_2\}$ (i.e., $d_2$ is the only distractor which is at least as salient as the target, $d_1$). The algorithm will select $\langle \mathsf{type}, \mathsf{man} \rangle$, which rules out the sole distractor $d_2$, leading to a successful reference ("The man"). If, however, $d_3$ would be the target then $C = \{d_1, d_2, d_3\} - \{d_3\} = \{d_1, d_2\}$, and the algorithm would operate as normal, producing a description realisable as "the man in the t-shirt". Krahmer and Theune chose to graft a variant of this idea onto the IA, but application to other algorithms is straightforward.

Krahmer and Theune (2002) compare two theories of computing linguistic salience – one based on the hierarchical focus constraints of Hajičová (1993), the other on the centering constraints of Grosz et al. (1995). They argue that the centering constraints, combined with a gradual decrease in salience of non-mentioned objects (as in the hierarchical focus approach) yields the most natural results. Interestingly, the need to compute salience scores can affect the architecture of the REG module. In Centering Theory, for instance, the salience of a referent is co-determined by the syntactic structure of the sentence in which the reference is realised; it matters whether the reference is in subject, object or another position. This suggests an architecture in which REG and syntactic realisation should be interleaved, a point to which we return below.

*Variations and extensions.* Differences in salience can be caused by *nonlinguistic* as well as linguistic factors: some domain objects may be further removed from the hearer than others, for example. Paraboni et al. (2007) demonstrated experimentally that such situations require substantial deviations from existing algorithms, to avoid causing unreasonable amounts of work to the reader. To see the idea, consider the way we refer to an address on a map: we probably don't say "Go to house number 3012 in Aberdeen", even if only one house in Aberdeen has that number; more likely we say something like "Go to house number 3012 So-and-so Road, in the West End of Aberdeen", adding logically redundant information specifically to aid the hearer's search.

Once referring expressions are no longer viewed as de-contextualised descriptions of their referent, a number of questions come up. When, for example, is it appropriate

to use a demonstrative ("this man", "that man"), or a pronoun ("he", "she")? As for demonstratives, it has proven remarkably difficult to decide when these should be used, and even harder to choose between the different types of demonstratives (e.g., Piwek (2008)). Concerning pronouns, Krahmer and Theune suggested that "he" abbreviates "the (most salient) man", and "she" "the (most salient) woman". In this way, algorithms for generating distinguishing descriptions might also become algorithms for pronoun generation. However, such an approach to pronoun generation is too simple, since additional factors are known to determine whether a pronoun is suitable or not (McCoy and Strube 1999; Henschel, Cheng, and Poesio 2000; Callaway and Lester 2002; Kibble and Power 2004). Based on analyses of naturally occurring texts, McCoy and Strube (1999), for example, emphasised the role of topics and discourse structure for pronoun generation, and pointed out that the changes in time scale are a reliable cue for this. In particular, they found that in certain places a definite description was used where a pronoun would have been unambiguous. This happened in particular when the time frame of the sentence differed from that of the sentence in which the previous mention occurred, as can be seen, for example, from a change in tense or a cue-phrase such as "several months ago". Kibble and Power (2004) use Centering Theory as their starting point in a constraint-based text generation framework, taking into account constraints such as salience, cohesion and continuity for the choice of referring expressions.

A lot of work on contextual reference takes *text* as its starting point (e.g., Poesio & Vieira (1998) and Belz et al. (2010)), unlike the majority of REG research, which uses standard knowledge representations of the kind exemplified in Table 1 (or some more sophisticated frameworks, see Section 4). An interesting variant is presented by Siddharthan and Copestake (2004), who set themselves the task of generating a referring expression at a specific point in a discourse, without assuming that a knowledge base (in the normal sense of the word) is available: all their algorithm has to go by is text. For example, a text might start saying "The new president applauded the old president". From this alone, the algorithm has to figure out whether, in the next sentence, it can talk about "the old president" (or some other suitable noun phrase) without risk of misinterpretation by the reader. The authors argue that standard REG methods can achieve reasonable results in such a setting, particularly (as we shall see next) with respect to the handling of lexical ambiguities that arise when a word can denote more than one property. Lexical issues such as these transcend the selection of semantic properties. Clearly, it is time for us to consider matters that lie beyond Content Determination.

### 3.5 Beyond Content Determination

In many early REG proposals, Lexical Choice and Surface Realisation follow Content Determination, in the style of a pipeline, with most of the actual research focussing predominantly on Content Determination. One might have thought that good results are easy to achieve by sending the output of the Content Determination module to a generic realiser (that is: a program converting meaning representations into natural language). With hindsight, any such expectations must probably count as naive.

Some REG studies have taken a different approach, interleaving Content Determination and Surface Realisation (Horacek 1997; Stone and Webber 1998; Krahmer and Theune 2002; Siddharthan and Copestake 2004), running counter to the pipeline architecture (Mellish et al. 2006). In this type of approach, syntactic structures are built up in tandem with semantic descriptions: when $\langle \mathsf{type}, \mathsf{man} \rangle$ has been added to the semantic description, a partial syntactic tree is constructed for a noun phrase, whose head noun is *man*. As more properties are added to the semantic description, appropriate modifiers

are slotted into the syntax tree; finally, the noun phrase is completed by choosing an appropriate determiner.

Even in these interleaved architectures, it is often assumed that there is a one-to-one correspondence between properties and words; but often a property can be expressed by different words, one of which may be more suitable than the other. Perhaps the most tangible reason why a word $w$ may be less ideal than another word $w'$ for expressing a property $p$ arises when $w$ is ambiguous between $p$ and some other property $p'$ (Siddharthan and Copestake 2004). One president may be "old" in the sense of *former*, while another is "old" in the sense of *aged*, in which case "the old president" can become ambiguous between the two people. To deal with the choice between "old" and "former", Siddharthan and Copestake propose to look at discriminatory power, the idea being that in this case "former" rules out more distractors than "old" (both presidents are old). One wonders, however, to what extent readers interpret ambiguous words "charitably": suppose *two* presidents are aged, while only *one* is the former president. In this situation, "the old president" seems clear enough, because only one of its two interpretations justifies the definite article (namely the one where "old" is to be understood as "former"). Clearly, people's processing of ambiguous expressions is an area where there is still a lot to explore.

If we turn away from Siddharthan and Copestake's setup, and return to the situation where generation starts from a non-textual knowledge base, similar problems with ambiguities may arise. In fact, the problem is not confined to Lexical Choice: ambiguities can arise during Surface Realisation as well. To see this, suppose Content Determination has selected the properties man and with telescope to refer to a person, and the result after Surface Realisation and Lexical Choice is "John saw *the man with the telescope*" then, once again, the clarity of the semantic description can be compromised by putting the description in a larger context, causing an attachment ambiguity, which may sometimes leave it unclear what man is the intended referent of the description. The generator can save the day by choosing a different realisation, generating "John saw *the man who holds the telescope*" instead. Similar ambiguities occur in conjoined references to plurals, as in "the old men and women", where "old" may or may not pertain to the women. These issues have been studied in some detail in connection with a small class of related coordination ambiguities (Chantree et al. 2005; Khan, van Deemter, and Ritchie 2008).

When the generated referring expressions are realised in a medium richer than plain text, for instance in the context of a virtual character (Gratch et al. 2002), another set of issues comes into play. It needs to be decided, then, which words should be emphasised in speech, possibly in combination with visual cues such as eyebrow movements and other gestures. Doing full justice to the expanding literature on multimodal reference is beyond the scope of this survey, but a few pointers may be useful. Various early studies looked at multimodal reference (Lester et al. 1999). One account, where pointing gestures directly enter the Content Determination module of REG, is presented by van der Sluis and Krahmer (2007), who focus on the trade-off between gestures and words. Kopp et al. (2008) are more ambitious, modelling different kinds of pointing gestures and integrating their approach with the generation strategy of Stone et al. (2003).

### 3.6 Discussion

Early REG research made a number of simplifying assumptions, and as a result the early REG algorithms could only generate a limited variety of referring expressions. When researchers started lifting some of these assumptions, this resulted in REG algorithms with an expanded repertoire, being able to generate, for instance, plural and relational

descriptions. However, this move created a number of new challenges. For instance, the number of ways in which one can refer to a set of target objects increases, so choosing a good one is more difficult as well. Should we prefer "the men *not* wearing an overcoat", "the young man *and* the old man" or "the men *left of* the woman". In addition, from a search perspective, the various proposals result in a larger search space, making computational issues more pressing. For some of the extensions (e.g., where boolean combinations of properties are concerned), the complexity of the resulting algorithm is substantially higher than that of the base IA. Moreover, researchers have often zoomed in on one extension of the IA, developing a new version which lifts one particular limitation. Combining all the different extensions into one algorithm which is capable of, say, generating references to salient sets of objects, using negations and relations and possibly vague properties, is a non-trivial enterprise. To give just one example, consider what happens when we combine salience with (other) gradable properties (cf., Sections 3.4 and 3.3). Should "the old man" be interpreted as 'the oldest of the men that are sufficiently salient' or 'the most salient of the men that are sufficiently old'? Expressions that combine gradable properties can easily become unclear, and determining when such combinations are nevertheless acceptable is an interesting challenge.

Some simplifying assumptions have only just begun to be lifted, through extensions that are only in their infancy, particularly in terms of their empirical validation. Other simplifying assumptions are still in place. For instance, there is a dearth of work that addresses functions of referring expressions other than mere identification. Similarly, even recent proposals tend to assume that it is unproblematic to determine what information is shared between speaker and hearer. We return to these issues in Section 6.

## 4. REG frameworks

Most early REG algorithms represent knowledge in a very basic way, specifically designed for REG. This may have been justified at the time, but years of research in Knowledge Representation (KR) suggest that such a carefree attitude towards the modelling of knowledge may not be wise in the long run. For example, when well-established KR frameworks are used, it may become possible to *re-use* existing algorithms for these frameworks, which have often been optimised for speed, and whose computational properties are well understood. Depending on the choice of framework, many other advantages can ensue. Since research that couples REG with KR is relatively new, and technical properties of the frameworks themselves can be easily found elsewhere, we shall be comparatively brief. For each framework, we focus on three questions: (a) How is domain information represented? (b) How is the semantic content of a referring expression represented? (c) How can distinguishing descriptions be found?

### 4.1 REG using Graph search

One of the first attempts to link REG with a more generic mathematical formalism was the proposal by Krahmer, van Erk and Verleg (2003), who used labelled directed graphs for this purpose. In this approach, objects are represented as the nodes (vertices) in a graph, and the properties of and relations between these objects are represented as edges connecting the nodes. Figure 4 shows a graph representation of our example domain. One-place relations (i.e., properties) such as man are modelled as loops (edges beginning and ending in the same node), while 2-place relations such as left of are modelled as edges between different nodes.

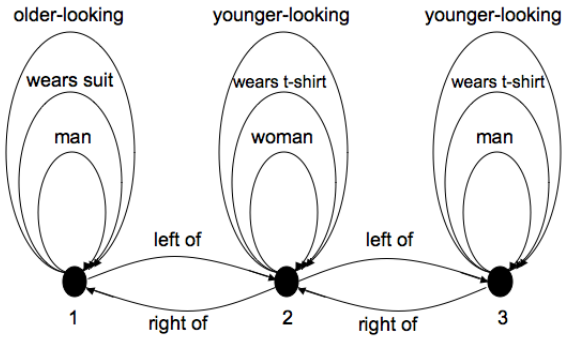


**Figure 4**
Representation of our example scene as a labelled directed graph.

Two kinds of graphs play a role: a scene graph representing the knowledge base, and referring graphs representing the content of referring expressions. The problem of finding a distinguishing referring expression can now be defined as a comparison between graphs. More specifically, it is a graph search problem: given a target object (i.e., a node in the scene graph), look for a distinguishing referring graph that is a subgraph of the scene graph and uniquely characterises the target. Intuitively, such a distinguishing graph can be "placed over" the target node with its associated edges, and not over any other node in the scene graph. The informal notion of one graph being "placed over" another corresponds with a subgraph isomorphism (Read and Corneil 1977). Figure 5 shows a number of referring graphs which can be placed over our target object $d_1$. The leftmost, which could be realised as "the man", fails to distinguish our target, since it can be "placed over" the scene graph in two different ways (over nodes 1 and 3).

Krahmer et al. (2003) use cost functions to guide the search process and to give preference to some solutions over others. They assume that these cost functions are monotonic, so extending a graph can never make it cheaper. Graphs are compatible with many different search algorithms, but Krahmer et al. (2003) employ a simple branch & bound algorithm for finding the cheapest distinguishing graph for a given target object. The algorithm starts from the graph containing only the node representing the target object and recursively tries to extend this graph by adding adjacent edges: edges starting from the target, or in any of the other vertices added later on to the referring graph under construction. For each referring graph, the algorithm checks which objects in the scene graph it may refer to, other than the target; these are the distractors. As soon as this set is empty, a distinguishing referring graph has been found. At this point, only alternatives that are cheaper than this best solution found so far need to be inspected. In the end, the algorithm returns the cheapest distinguishing graph which refers to the target, if one exists, otherwise it returns the empty graph.

One way to define the cost function would be to assign each edge a cost of one point. Then the algorithm will output the smallest graph that distinguishes a target (if one exists), just as the Full Brevity algorithm would. Alternatively, one could assign costs in accordance with the list of preferred attributes in the IA, making more preferred properties cheaper than less preferred ones. A third possibility is to compute the costs of an edge $e$ in terms of the probability $P(e)$ that $e$ occurs in a distinguishing description (which can be estimated by counting occurrences in a corpus), making frequent properties cheap and rare ones expensive:

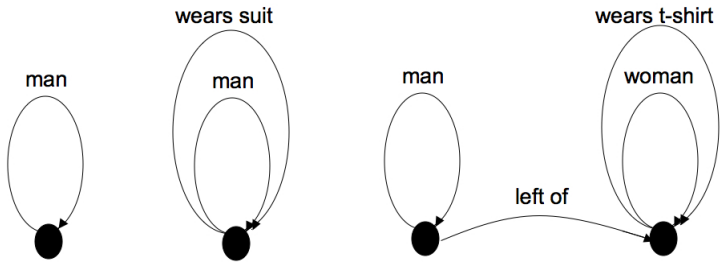


**Figure 5**
Some referring graphs for target $d_1$.

$$cost(e) = -\log_2(P(e))$$

Experiments with stochastic cost functions have shown that these enable the graph-based algorithm to capture a lot of the flexibility of human references (Krahmer et al. 2008; Viethen et al. 2008).

In the graph-based perspective, relations are treated in the same way as individual properties, and there is no risk of running into infinite loops ("the cup to the left of the saucer to the right of the cup ..."). Unlike Dale and Haddock (1991) and Kelleher and Kruijff (2006), no special measures are required, since a relational edge is either included in a referring graph or not: including it twice is not possible. Van Deemter and Krahmer (2007) show that many of the proposals discussed in Section 3 can be recast in terms of graphs. They argue, however, that the graph-based approach is ill-suited for representing disjunctive information. Here, the fact that directed graphs are not a fully fledged KR formalism makes itself felt. Whenever a REG algorithm needs to reason with complex information, heavier machinery is required.

### 4.2 REG using Constraint satisfaction

Constraint satisfaction is a computational paradigm that allows efficient solving of NP hard combinatoric problems such as scheduling (van Hentenryck 1989). It is among the earliest frameworks proposed for REG (Dale and Haddock 1991), but in later years, this approach has seldom been emphasised (with notable exceptions, such as Stone and Webber (1998)), until Gardent (2002) showed how constraint programming can be used to generate expressions that refer to sets. She proposed to represent a description $L$ for a target set $S$ as a pair of set variables:

$$L_S = \langle P_S^+, P_S^- \rangle,$$

where one variable ($P_S^+$) ranges over sets of properties that are *true* of the elements in $S$ and the other ($P_S^-$) over properties that are *false* of the elements in $S$. The challenge – taken care of by existing constraint solving programs – is to find suitable values (i.e., sets of properties) for these variables. To be "suitable", values need to fulfil a number of REG-style constraints:

1. All the properties in $P_S^+$ are true of all elements in $S$.
2. All the properties in $P_S^-$ are false of all elements in $S$.

3. For each distractor $d$ there is a property in $P_S^+$ which is false of $d$, or there is a property in $P_S^-$ which is true of $d$.

The third clause says that every distractor is ruled out by either a positive property (i.e., a property in $P_S^+$) or a negative property (i.e., a property in $P_S^-$), or both. An example of a distinguishing description for the singleton target set $\{d_1\}$ in our example scene would be $\langle\{\mathsf{man}\}, \{\mathsf{right}\}\rangle$, since $d_1$ is the only object in the domain who is both a man and not on the right. The approach can be adapted to accommodate disjunctive properties to enable reference to sets (Gardent 2002).

Constraint satisfaction is compatible with a variety of search strategies (Kumar 1992). Gardent opts for a "propagate-and-distribute" strategy, which means that solutions are searched for in increasing size, first looking for single properties, next for combinations of two properties, etc. This amounts to the Full Brevity search strategy, of course. Accordingly, Gardent's algorithm yields a minimal distinguishing description for a target, provided one exists. Given the empirical questions associated with Full Brevity, it may well be worthwhile to explore alternative search strategies.

The constraint approach allows an elegant separation between the specification of the REG problem and its implementation. Moreover, the handling of relations is straightforwardly applicable to relations with arbitrary numbers of arguments. Gardent's approach does not run into the aforementioned problems with infinite loops, because a set of properties (being a set) cannot contain duplicates. Yet, like the labelled graphs, the approach proposed by Gardent has significant limitations, which stem from the fact that it does not rest on a fully developed KR system. General axioms cannot be expressed, let alone enter logical deduction. We are forced to re-visit the question of what is the best way for REG to represent and reason with knowledge.

### 4.3 REG using modern Knowledge Representation

To find out what is missing, let us see what happens when domains scale up. Consider a furniture domain, and suppose every chair is in a room, that every room is in an apartment, and every apartment in a house. Listing all relevant relations between individual objects separately ("chair $a$ is in room $b$", "room $b$ is in apartment $c$", "chair $a$ is in apartment $c$", "apartment $c$ is in house $d$") is onerous, error prone, space-consuming and messy. Modern KR systems solve this problem by employing general *axioms* (e.g., expressing transitivity of the "in" relation; if $x$ is in $y$, and $y$ is in $z$, then $x$ is in $z$). Logical *inference* allows the KR system to derive implicit information. For example, from "chair $a$ is in room $b$", "room $b$ is in apartment $c$", and "apartment $c$ is in house $d$", the transitivity of "in" allows us to infer that "chair $a$ is in house $d$". This combination of basic facts and general axioms allows a succinct and insightful representation of facts.

Modern KR comes in different flavours. Recently, two different KR frameworks have been linked with REG, one based on Conceptual Graphs (Croitoru and van Deemter 2007), the other on Description Logics (Gardent and Striegnitz 2007; Areces, Koller, and Striegnitz 2008). The first have their origin in Sowa (1984) and were greatly enhanced by Baget and Mugnier (2002). The latter grew out of work on KL-ONE (Brachman and Schmolze 1985) and became even more prominent in the wider world of computing when they came to be linked with the ontology language OWL, which underpins current work on the semantic web (Baader et al. 2003). Both formalisms represent attempts to carve out computationally tractable fragments of First-Order

Predicate Logic for defining and reasoning about concepts, and are closely related (Kerdiles 2001). For reasons of space, we focus on Description Logic.

The basic idea is that a referring expression can be modelled as a formula of DL, and that REG can be viewed as the problem of finding a formula that denotes (i.e., refers to) the target set of individuals. Let us revisit our example domain, casting it as logical model $M$, as follows: $M = \langle D, \|.\| \rangle$, where $D$ (the domain) is a finite set $\{d_1, d_2, d_3\}$ and $\|.\|$ is an interpretation function which gives the denotation of the relevant predicates (thus: $\|\mathsf{man}\| = \{d_1, d_3\}$, $\|\mathsf{left\text{-}of}\| = \{\langle d_1, d_2\rangle, \langle d_2, d_3\rangle\}$ etc.). Now the REG task can be formalised as: given a model $M$ and a target set $S \subseteq D$, look for a Description Logic formula $\varphi$ such that $\|\varphi\| = S$. The following three expressions are the Description Logic counterparts of the referring graphs in Figure 5:

(a) $\mathsf{man}$
(b) $\mathsf{man} \sqcap \mathsf{wears\ suit}$
(c) $\mathsf{man} \sqcap \exists\ \mathsf{left\text{-}of}.(\mathsf{woman} \sqcap \mathsf{wears\ t\text{-}shirt})$

The first, (a), would not be distinguishing for $d_1$ (since its denotation includes $d_3$), but (b) and (c) would. Note that $\sqcap$ represents the conjunction of properties, and $\exists$ represents existential restriction. Negations can be added straightforwardly, as in $\mathsf{man} \sqcap \neg\ \mathsf{wears\ suit}$, which denotes $d_3$.

Areces et al. (2008) search for referring expressions in a somewhat non-standard way. In particular, their algorithm does not start with one particular target referent: it simply attempts to find the different sets that can be referred to. They start from the observation that REG can be reduced to computing the similarity set of each domain object. The similarity set of an individual $x$ is the set of those individuals that have all the properties that $x$ has. Areces et al. (2008) present an algorithm (based on a proposal by Hopcroft (1971)) which computes the similarity sets, along with a DL formula associated with each set. The algorithm starts by partitioning the domain using atomic concepts such as man and woman, which splits the domain in two subsets ($\{d_1, d_3\}$ and $\{d_2\}$ respectively). At the next stage, finer partitions are made by making use of concepts of the form $\exists R.AtomicConcept$ (e.g., men left of a woman), and so on, always using concepts established during one phase to construct more complex concepts during the next. All objects are considered in parallel, so there is no risk of infinite loops. Control over the output formulae is achieved by specifying an incremental preference order over possible expressions, but alternative control strategies could have been chosen.

### 4.4 Discussion

Even though the role of KR frameworks for REG has received a fair bit of attention in recent years, one can argue that this constitutes just the first steps of a longer journey. The question of which KR framework suits REG best, for example, is still open; which framework has the best coverage, which allows all useful descriptions to be expressed? Moreover, can referring expressions be found quickly in a given framework, and is it feasible to convert these representations into adequate linguistic realisations? Given the wealth of possibilities offered by these frameworks, it is remarkable that much of their potential is often left unused. In Areces et al.'s proposal, for example, generic axioms do not play a role, nor does logical inference. Ren, van Deemter and Pan (2010) sketch how REG can benefit if the full power of KR is brought to bear, using DL as an example. They show how generic axioms can be exploited, as in the example of the

furniture domain, where a simple transitivity axiom (if $x$ is in $y$, and $y$ is in $z$, then $x$ is in $z$) allows a more succinct and insightful representation of knowledge. Similarly, *incomplete* information can be used, as when we know that someone is either Dutch or Belgian, without knowing which of the two. Finally, by making use of more expressive DL fragments, it becomes possible to identify objects that previous REG algorithms were unable to identify, as when we say "the man who owns three dogs", or "the man who only kisses women", referring expressions that were typically not considered by previous REG algorithms.

Extensions of this kind raise new empricial questions, as well. It is an open question, for instance, when human speakers would be inclined to use such complex descriptions. These problems existed even in the days of the classic REG algorithms (when it was already possible to generate lengthy descriptions) but they have become more acute now that it is possible to generate *structurally* complex expressions as well. There is a clear need for empirical work here, which might teach us how the power of these formalisms ought to be constrained.

## 5. Evaluating REG

In the time up to and including (Dale and Reiter 1995), evaluation of REG algorithms received virtually no attention. More recently, evaluation studies have started to be carried out more and more often. Most of these were predicated on the (often implicit) assumption – which we shall debate in section 7 – that REG algorithms should try to generate expressions that are optimally similar to these produced by human speakers or writers. The dominant method at the moment is, accordingly, to measure the similarity between generated expressions and the ones in a suitable corpus of referring expressions. REG came late to corpus-based evaluation (compared to other parts of computational linguistics) because suitable data sets are hard to come by. In this section, we discuss what criteria a data set should meet to make it suitable for REG evaluation, and we survey which collections are currently available. In addition, we discuss how one is to determine the performance of a REG algorithm on a given data set. As we shall see, a lot of work has been done in recent years, but there are still significant questions, particularly regarding the relation between automatic metrics and human judgements.

### 5.1 Corpora for REG evaluation

Text corpora are full of referring expressions. For evaluating the *realisation* of referring expressions, such corpora are very suitable, and various researchers have used them, for instance to evaluate algorithms for modifier orderings (Shaw and Hatzivassiloglou 1999; Malouf 2000; Mitchell 2009). Text corpora are also important for the study of anaphoric links between referring expressions. The texts that make up the GNOME corpus (Poesio et al. 2004), for instance, contain descriptions of museum objects and medical patient information leaflets, with each of the two subcorpora containing some 6000 NPs. A lot of information is marked up, including anaphoric links. Yet, text corpora of this kind are of limited value for evaluating the content selection part of REG algorithms. For that, one needs a corpus that is fully "semantically transparent" (van Deemter, van der Sluis, and Gatt 2006): a corpus that contains the actual properties of all domain objects as well as the properties that were selected for inclusion in a given reference to the target. Text corpora such as GNOME do not meet this requirement, and it is often difficult or impossible to add all necessary information, because of the size and complexity of the relevant domains. For this reason, data sets for content

selection evaluation are typically collected via experiments with human participants in simple and controlled settings. Broadly speaking, two kinds of experimental corpora can be distinguished: corpora specifically collected with reference in mind, and corpora collected wholly or partly for other purposes, but which have nevertheless been analysed for the referring expressions in them. We will briefly sketch some corpora of the latter kind, after which we shall discuss the former in more detail.

*General-purpose corpora*. One way to elicit "natural" references is to let participants perform a task for which they need to refer to objects. An example is the corpus of so-called *pear stories* of Chafe (1980), in which people were asked to describe a movie about a man harvesting pears, in a fluent narrative. The resulting narratives featured such sequences as "And he fills his thing with pears, and comes down and there's a basket he puts them in. (...) And then a boy comes by, on a bicycle, the man is in the tree, and the boy gets off his bicycle (...)", where a limited set of individuals come up several times. The referring expressions in a subset of these stories were analysed in Passonneau (1996), who asked how the form of the re-descriptions (such as "he", "them", and "the man") in these narratives might best be predicted, comparing "informational" considerations (which form the core of most algorithms in the tradition started by Dale and Reiter, as we have seen) with considerations based on Centering Theory (Grosz, Joshi, and Weinstein 1995). Passonneau, who tested her rules on 319 noun phrases, found support for an integrated model, where centering constraints take precedence over informational considerations.

The well-known MapTask corpus (Anderson et al. 1991) is another example of a corpus in which reference plays an important role. It consists of dialogues between two participants; both have maps with landmarks indicated, but only one (the instruction giver) has a route on the map and he or she instructs the other (the follower) about this particular route. Referring expressions are routinely produced in this task to refer to the landmarks on the maps ("the cliff"). Participants use these not only for identification purposes but also, for instance, to verify whether they understood their dialogue partner correctly. In the original MapTask corpus, the landmarks were labeled with proper names ("cliff"), making them less suitable for studying content determination. To facilitate the study of reference, the iMap corpus was created (Guhe and Bard 2008), a modified version of the MapTask corpus where landmarks are not labelled, and systematically differ along a number of dimensions, including type (e.g., owl, penguin, etc.), number (singular, plural) and colour; a target may thus be referred to as "the two purple owls". Since participants may refer to targets more than once, it becomes possible to study initial and subsequent reference (Viethen et al. 2010).

Yet another example is the Coconut corpus (Di Eugenio et al. 2000), a set of task-oriented dialogues in which participants negotiate which furniture items they want to buy on a fixed, shared budget. Referring expressions in this corpus ("a yellow rug for 150 dollars") do not only contain information to identify a particular piece of furniture, but also include properties which directly refer to the task at hand (e.g., how much money is still available for a particular furniture item and what the state of agreement between the negotiators is).

An attractive aspect of these corpora is that they represent fairly realistic communication, related to a more or less natural task. However, in these corpora, the identification of objects tends to be mixed with other communicative tasks (verification, negotiating). This does not mean that the corpora in question are unsuitable for the study of reference, of course. More specifically, they have been used for evaluating REG algorithms, to compare the performance of traditional algorithms

**Table 2**
Overview of dedicated Referring Expression corpora (alphabetical), with for each corpus an indication of the domain, and the number of participants and collected descriptions.

| Corpus Name | Reference | Domain | Participants | Descriptions |
|---|---|---|---|---|
| Bishop | Gorniak & Roy (2004) | Coloured cones in 3D scene | 9 | 447 |
| Drawer | Viethen & Dale (2006) | Drawers in filing cabinet | ? | 160 |
| GRE3D3 | Viethen & Dale (2008) | Spheres, Cubes in 3D scene | 63 | 630 |
| iMap | Guhe & Bard (2008) | Various objects on a map | 64 | 9567 |
| TUNA | van Deemter et al. (2011) | Furniture, People | 60 | 2280 |

with special-purpose algorithms that take dialogue context into account (Passonneau 1996; Jordan and Walker 2005; Gupta and Stent 2005). For example, when the speaker attempts to persuade the hearer to buy an item, Jordan's *Intentional Influences* algorithm selects those properties of the item that make it a better solution than a previously discussed item. In yet other situations – for example, when a summarisation is offered – *all* mutually known properties of the item are selected. Jordan's algorithm outperforms traditional algorithms, which is not surprising given that the latter were not designed to deal with references in interactive settings (Jordan 2000).

*Dedicated corpora.* In recent years, a number of new corpora have been collected, specifically focussing on the types of referring expressions that we are focussing on in this survey. A number of such corpora are summarised in Table 2. In some ways, these corpora are remarkably similar. Reflecting the prevalent aims of research on REG, for example, they focus on expressions that aim to identify their referent "in one shot", disregarding the linguistic context of the expression, i.e. in the "null context", as it is sometimes called (Viethen and Dale 2007). In all these corpora, participants were asked to refer to targets in a visual scene also containing the distractors. This setup means that the properties of target objects and their distractors are known, which makes it comparatively easy to make these corpora semantically transparent by annotating the references that were produced. In addition, most corpora are "pragmatically transparent" as well, meaning that the communicative goals of the participants were known (typically identification).

An early example is the Bishop corpus (Gorniak and Roy 2004). For this data set, participants were asked to describe objects in various computer generated scenes. Each of these scenes contained up to 30 objects ("cones") randomly positioned on a virtual surface. All objects had the same shape and size, and hence targets could only be distinguished using their colour (either green or purple) and their location on the surface ("the green cone at the left bottom"). Each participant was asked to identity targets in one shot, and for the benefit of an addressee who was physically present but did not interact with the participant.

The Drawer corpus, collected by Viethen and Dale (2006), has a similar objective, but here targets are real, being one of 16 coloured drawers in a filing cabinet. On different occasions, participants were given a random number between 1 and 16 and asked to refer to the corresponding drawer for an onlooker. Naturally, they were asked not to use the number; instead they could refer to the target drawers using colour, row and column, or some combination of those. In this corpus, referring expressions ("the pink

drawer in the first row, third column") once again solely serve an identification purpose. Viethen and Dale (2008) collected a new corpus (GRE3D3), specifically looking at when participants use spatial relations. For this data collection, participants were presented with 3D scenes (made with Google SketchUp) containing three simple geometric objects (spheres and cubes of different colours and sizes, and in different configurations), of which one was the target. Viethen and Dale (2008) found that spatial relations were frequently used ("the ball in front of the cube"), even though they were never required for identification. Whether this generalises to other visual scenes (in which spatial relations are less immediately 'available') is an interesting question for future research.

The TUNA corpus (Gatt, van der Sluis, and van Deemter 2007; van Deemter et al. 2011) was collected via a web-based experiment, in which singular and plural descriptions were gathered by showing participants one or two targets, where the plural targets could either be similar (same type) or dissimilar (different type). Targets were always displayed with 6 distractors, and the resulting domain objects were randomly positioned in a 3 x 5 grid, with targets surrounded by a red border. Example trials are shown in Figure 6.

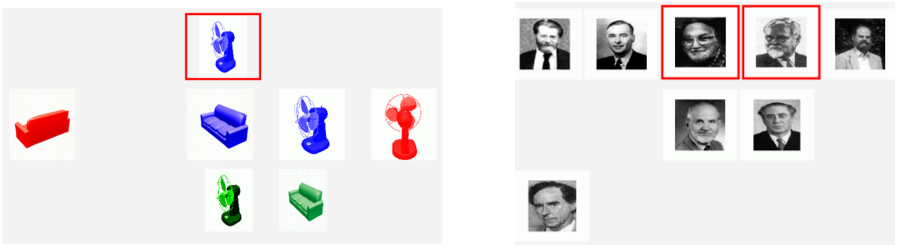

**Figure 6**
Example trials from the TUNA corpus, a singular trial for the furniture domain ("the small blue fan", left) and a plural trial for the people domain ("the men with glasses", right).

The corpus contains two different domains: a furniture and a people domain. The first domain is based on pictures of furniture and household items, taken from the Object Databank (produced by Michael Tarr's lab, see `http://www.tarrlab.org/`). These were manipulated so that besides type (chair, desk, fan) also colour, orientation and size could systematically be varied. The number of possible attributes and values in the people domain is much larger (and more difficult to pin down); this domain consists of a set of black and white photographs of people (all famous mathematicians) used in an earlier study of van der Sluis and Krahmer (2007). Properties of these photographs include gender, head orientation, age, beard, hair, glasses, suit, shirt and tie. It is interesting to realize that the TUNA corpus was designed to have one shortest description for each target, while in other data sets, such as Viethen and Dale's (2006) drawer corpus, a single shortest description does not always exist. The TUNA corpus has formed the basis of three shared REG challenges, to which we turn below.

### 5.2 Evaluation metrics

How to compare human references with those produced by a REG algorithm? When looking for measures that compute the content overlap, one source of inspiration may come from biology and information retrieval (van Rijsbergen 1979). One measure used in these fields is the Dice (1945) coefficient, which was originally proposed to quantify ecologic association between species, and was first applied to REG by Gatt et al. (2007).

The Dice coefficient (which is not dissimilar to the "match" function used by Jordan (2000)) is computed by scaling the number of elements that two sets have in common, by the size of the two sets combined:

$$\mathit{Dice}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

The Dice measure ranges from 0 (no agreement, i.e., no elements shared between A and B) to 1 (complete agreement; A and B share all elements). For REG, $A$ and $B$ can be understood as attributes (e.g., type) or as attribute–value pairs (properties; $\langle \mathsf{type}, \mathsf{man} \rangle$). The former option tends to be used in earlier work, but has the somewhat counterintuitive consequence that two descriptions which express different values of the same attribute ("the man" and "the woman", say, or "the dog" and "the chihuahua", in the earlier discussed cats-and-dogs example) have a Dice score of 1. Hence, in the discussion below we shall measure overlap in terms of properties.

An alternative to Dice that is sometimes used is the MASI (Measuring Agreement on Set-valued Items) metric of Passonneau (2006):

$$\mathit{MASI}(A, B) = \delta \times \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

This is basically an extension of the well-known Jaccard (1901) metric with a weighting function $\delta$ which biases the score in favour of similarity where one set is a sub- or a superset of the other:

$$\delta = \begin{cases} 1, & \text{if } A = B \\ \frac{2}{3}, & \text{if } A \subset B \text{ or } B \subset A \\ 0, & \text{if } A \cap B = \emptyset \\ \frac{1}{3}, & \text{otherwise} \end{cases} \tag{3}$$

Dice and MASI are straightforward measure for overlap, but they do have their disadvantages. For example, they assume that all properties are independent and that all are equally different from each other. Suppose a human participant referred to $d_1$ in our example domain as "the man in the suit next to a woman", and consider the following two references produced by a REG algorithm: "the man in the suit" and "the man next to a woman". Both omit one property from the human reference and thus have the same Dice and MASI scores. But only the former reference is distinguishing; the latter is not. This problem could be solved, for example, by adopting a binary weighted version of the metrics which multiply the resulting score with 1 for a distinguishing description and with 0 for a non-distinguishing one.

A more general issue with these overlap metrics can be illustrated with an example from Richard Power (p.c.). Consider the two (roughly equivalent) expressions "the palomino" and "the horse with the gold coat and white mane and tail". Straightforward counting of attribute–value pairs would result in an overlap score of zero, which would be misleading, since the two descriptions express essentially the same content, with the latter description combining, in one property, all properties expressed in the former.

This problem clearly calls for a more principled approach to representing and counting properties.

During evaluations, Dice or MASI scores are typically averaged over references for different trials and produced by different human participants, making them fairly rough measures. It could be that an algorithm's predictions match the descriptions of some participants very well, but those of other participants not at all. To partially compensate for this, sometimes also the proportion of times an algorithm achieves a perfect match with a human reference is reported. This measure is known, somewhat confusingly, as Recall (Viethen and Dale 2006), the Perfect Recall Percentage (PRP) (Gatt, van der Sluis, and van Deemter 2007) and Accuracy (Gatt, Belz, and Kow 2008).

The measures discussed so far do not take the actual linguistic realisation of the referring expressions into account. For these, string distance metrics are obvious candidates, since these have proven their worth in various other areas of computational linguistics. One well-known string distance metric, which has also been proposed for REG evaluation, is the Levenshtein (1966) distance: the minimal number of insertions, deletions and substitutions needed to convert one string into another, possibly normalised with respect to length (Bangalore, Rambow, and Whittaker 2000). The BLEU (Papineni et al. 2002) and NIST (Doddington 2002) metrics have their origin in machine translation evaluation . BLEU measures $n$-gram overlap between strings; for machine translation $n$ is often set to 4, but given that referring expressions tend to be short, $n = 3$ seems a better option for REG evaluation (Gatt, Belz, and Kow 2009). NIST is a BLEU variant giving more importance to less frequent (and hence more informative) $n$-grams. Finally, Belz and Gatt (2008) also use the rouge-2 and rouge-su4 measures (Lin and Hovy 2003), originally proposed for evaluating automatically generated summaries.

An obvious benefit of these string metrics is that they are easy to compute automatically, while property-based evaluation measures such as Dice require an extensive manual annotation of selected properties. However, the added value of string-based metrics for REG is relatively unclear. It is not obvious, for instance, that a smaller Levenshtein distance is always to be preferred over a longer one; the expressions "the man wearing a t-shirt" and "the woman wearing a t-shirt" are at a mere 2 Levenshtein distance from each other, but only the former would be a good description for target $d_3$. On the other hand, "the male person on the right" is at a Levenshtein distance of 15 from "the man wearing a t-shirt", and both are perfect descriptions of $d_3$.

Alternatively, referring expressions could also be evaluated by human judges, although this obviously is more time consuming than an automatic evaluation. Gatt et al. (2009) collected judgements of *Adequacy* ("How clear is this description? Try to imagine someone who could see the same grid with the same pictures, but didn't know which of the pictures was the target. How easily would they be able to find it, based on the phrase given?") and *Fluency* ("How fluent is this description? (...) Is it good, clear English?"). One may also be interested in the extent to which references are useful for addressees. This can be evaluated in a number of different ways. Belz and Gatt (2008), for example, first show participants a generated description for a trial. After participants have read this description, a scene appears and participants are asked to click on the intended target. This yields three extrinsic evaluation metrics: the reading time, the identification time and the error rate, defined as the number of incorrectly identified targets.

### 5.3 Discussion

Three lessons can be learnt from the recent work on evaluation. First, the emergence of transparent corpora after 1995 has greatly facilitated the empirical evaluation of REG al-

gorithms, particularly for content selection. Focussing on reference in simple situations, a number of studies based on transparent corpora found that the IA outperformed the Full Brevity and Greedy Heuristic algorithms (Viethen and Dale 2006; van Deemter et al. 2011). There is an important catch, however: as demonstrated in van Deemter et al. (2011), the performance of the IA crucially depends on the chosen preference order: the best preference order outperforms the other two algorithms, but many other preference orders perform far worse. This is a problem, since no procedure for finding a good preference order is known. (For $n$ attributes, there are $n!$ preference orders to consider, so trial and error is not an option except in extremely simple cases.) Perhaps most controversially, the authors argue that the evidence is starting to stack up in favour of the thesis that the Greedy algorithm – or variants of the Greedy algorithm that choose properties on the basis of more than just their discriminatory power – might be superior to algorithms that use the same preference order all the time.

Second, evaluations suggest that human-produced referring expressions differ from automatically generated references in a number of ways. Human references often include redundant information, making the references overspecified in ways that were not accounted for by standard REG algorithms. An additional problem is that there appears to be considerable individual variation, both within and between speakers, which is something that existing REG algorithms do not model (Dale and Viethen 2010).

Third, it is still somewhat unclear what the best REG evaluation metrics are. The three REG Challenges based on the TUNA set-up offer a wealth of information in this respect (Gatt and Belz 2010). In each of these challenges, a number of research teams submitted one or more REG generation systems, allowing detailed statistical analyses over the various metrics. It was found that Dice, MASI and PRP are very highly correlated (all $r > .95$). Interestingly, these metrics correlate negatively with the proportion of references that are minimally specified (Gatt, Belz, and Kow 2008); in other words, systems that produce more overspecified references tend to do better in terms of Dice and other overlap metrics. Concerning the surface realisation metrics, it was found that – when comparing different realisations of a given set of attributes – the NIST and BLEU string metrics correlate strongly with each other ($r = .9$), as one might expect, but neither correlates well with Levenshtein distance (Gatt, Belz, and Kow 2008).

As for the extrinsic measures, Gatt et al. (2008) only report a significant correlation between reading time and identification time, which suggests that slow readers are also slow identifiers, or that referring expressions that are hard to read also make it harder to identify the intended referent. Gatt et al. (2009) let participants *listen* to expressions that were produced either automatically or by human speakers, and found a strong correlation between identification accuracy and adequacy, suggesting that more adequate references also have more correct identifications. Also, they found a negative correlation between fluency and identification time, implying that more fluent descriptions reduce the identification time.

It is notable that essentially no correlations were found between these extrinsic task performance measures and the automatic metrics for human-likeness (Belz and Gatt 2008; Gatt and Belz 2010). Different explanations are possible for this lack of a correlation. Gatt and Belz (2010), in discussing this issue, note that the nature of the TUNA data could be partly responsible. The TUNA data collection was done in a web-based and relatively unrestricted manner, and idiosyncratic references do occur in it ("a red chair, if you sit on it, your feet would show the south east"). It is therefore possible that a better corpus would show up a correlation between the two kinds of metrics. Alternatively, it could be that people are not always very good at designing their utterances in a way that is optimal for hearers (Horton and Keysar 1996) (see also

Section 6), so producing descriptions that resemble human-produced ones is not the same as producing descriptions that are of optimal use for hearers. This suggests that the two sets of metrics measure different things, and that they correspond with two different aims that the designer of an REG algorithm might have: one set of metrics could be used if the aim is to mimic speakers, another if the aim is to produce optimal benefits for hearers.

So far, experimental evaluation has mostly been limited to the simplest of situations, focussing on algorithms that produce singular descriptions, expressing conjunctions of basic properties in small and artificial domains. Most of the extensions discussed in Section 3 have not been evaluated systematically. Moreover, tasks such as the one on which the TUNA corpus is based can be argued not to be "ecologically valid": human participants produce type-written expressions for an imaginary audience on the basis of abstract visual scenes. The effects of these limitations on the descriptions produced are partly unknown, although some re-assuring results have recently been obtained. It has been shown, for example, that, speakers who address an imaginary audience refer in similar ways to those who address an audience that is physically present (van der Wege 2009). Similarly, Koolen et al. (2009) show that speaking rather than typing has no effect on the kind and number of attributes in the referring expressions that are produced, although speakers tend to use more words than typists to convey the same amount of information. It would be valuable to evaluate REG algorithms in the context of a specific application, so the added value of different REG algorithms for a real-life application can be gauged (Gatt, Belz, and Kow 2009).

Two recent evaluation challenges seem promising for these reasons. GREC (Belz et al. 2010) focusses on the task of deciding which form a referring expression should take in a textual context, which is important for generating coherent texts such as summaries (see also Section 6). GIVE (Koller et al. 2010) focusses on generating directions in a virtual 3D environment, where reference is only one task among a number of others. This new challenge has so far not included a separate test of REG algorithms employed in the systems submitted, but it seems likely that GIVE will cause REG research to focus on harder tasks, including reference in discourse context, reference to sets, and references that are spread out over several utterances (e.g., Denis (2010)).

## 6. Open issues

In the previous sections we have discussed three main dimensions in which REG research has moved beyond the state-of-the-art of 2000. Along the way, various loose ends have been identified. For example, not all simplifying assumptions of early REG work have been adequately addressed, and the enterprise of combining extensions is still in its infancy (Section 3). It is still unclear whether referring expressions in advanced knowledge representation frameworks can be found quickly (Section 4), and empirical data has only been collected for the simplest referring expressions (Section 5). In this section, we suggest six further questions for future research.

**1. How to match a REG algorithm to a particular domain and application?** Evaluation of classic REG algorithms has shown that with some preference orders, the IA outperformed the Full Brevity and Greedy Heuristic algorithms, but with others it performed much worse than these (van Deemter et al. 2011). The point is that the IA, as it stands, is under-determined, because it does not contain a procedure for finding a preference order. Sometimes psycholinguistic experiments come to our aid, for instance Pechmann's (1989) study showing that speakers have a preference for absolute

properties (colour) over relative ones (size). Unfortunately, for most other attributes, no such experiments have been done.

It seems reasonable to assume that frequency tells us something about preference: a property that is used frequently is also more likely to be high on the list of preferred properties (Gatt and Belz 2010; van Deemter et al. 2011). But suitable corpora to determine preferences are rare, as we have seen, and their construction is time consuming. This raises the question how much data would be needed to make reasonable guesses about preferred properties; this could be studied, for instance, by drawing learning curves where increasingly large proportions of a transparent corpus are used to estimate a preference order and the corresponding performance is measured.

The IA is more drastically under-determined than most other algorithms: the Full Brevity and the Greedy Heuristic algorithm are specified completely up to situations where there is a tie: a tie between two equally lengthy descriptions in the first case, and a tie between two properties that have the same discriminatory power in the second. To resolve such ties frequency data would clearly be helpful. Similar questions apply to other generation algorithms. For instance, the graph-based algorithm as described by Krahmer et al. (2008) assigns one of three different costs to properties (they can be free, cheap, or somewhat expensive), and frequency data is used to determine which costs should be assigned to which properties (properties that are almost always used in a particular domain can be for free, etc.). A recent experiment (Theune et al. 2011) suggests that training the graph-based algorithm on a corpus with a few dozen items may already lead to a good performance. In general, knowing how much data is required for a new domain to reach a good level of performance is an important open problem for many REG algorithms.

**2. How to move beyond the "paradigms" of reference?** A substantial amount of REG research focusses on what we referred to in the Introduction as the "paradigms" of reference: "first-mention" distinguishing descriptions consisting of a noun phrase starting with "the", which serve to identify some target, and which do so without any further context. But how frequent are these "paradigmatic" kinds of referring expressions? Poesio and Vieira (1998), in one of the few systematic attempts to quantify the frequency of different uses of definite descriptions in segments of the Wall Street Journal corpus, reported that "first mention definite descriptions" are indeed the most frequent in these texts. These descriptions often do not refer to visual objects in terms of perceptual properties but to more abstract entities. One might think that it matters little whether a description refers to a perceivable object or not; a description like "the third quarter" rules out three quarters much like "the younger-looking man" in our example scene rules out the older-looking distractor. It appears, however, that the representation of the relevant facts in such cases tends to be a more complicated affair, and it is here particularly that more advanced knowledge representation formalisms of the kind discussed in Section 4 come into their own (a point to which we return below).

Even though first-mention definite descriptions are the most frequent in Poesio and Vieira's sample, other uses abound, including anaphoric descriptions and bridging descriptions, whose generation is studied by Gardent and Striegnitz (2007). Pronouns come to mind as well. The content determination problem for these other kinds of referring expressions may not be overly complex, but deciding where in a text or dialogue each kind of referring expression should be used is hard. Still, this is an important issue for, for example, automatic summarisation. One of the problems of extractive summaries is that co-reference chains may be broken, resulting in less coherent texts. Regeneration of referring expressions is a potentially attractive way of regaining some

of the coherence of the source document (Nenkova and McKeown 2003). Finally, there are proper names. REG research often works from the assumption that referents can not be identified through proper names. (If proper names were allowed, why bother inventing a description?) But in real text, proper names are highly frequent. This does not only raise the question when it's best to use a proper name, or which version of a proper name should be used (is it "Prince Andrei Nikolayevich Bolkonsky", "Andrei Bolkonsky", or just "Andrei"?), but also how proper names can occur as part of a larger description, as when we refer to a person using the description "the author of Primary Colours", for example, where the proper name "Primary Colours" refers to a well known book (whose author was long unknown). Surely, it is time for REG to turn proper names into first-class citizens.

Generation of referring expressions in a text is studied in the GREC (Generating Referring Expressions in Context) challenges (Belz et al. 2008). A corpus of wikipedia texts (for cities, countries, rivers, persons and mountains) was constructed, and in each text all elements of the coreference chain for the main subject were removed. For each of the resulting reference gaps, a list of alternative referring expressions, referring to the subject, was given (including the "correct" reference, i.e., the one that was removed from the text). One well-performing entry (Hendrickx et al. 2008) predicted the correct type of referring expression in 76% of the cases, using a memory-based learner. These results suggest that it is feasible to learn which type of referring expression is best in which instance. If so, REG in context could be conceived of as a two-stage procedure where first the form of a reference is predicted, after which the content and realisation are determined. REG algorithms as described in the present survey would naturally fit into this second phase. It would be interesting to see if such a method could be developed for a data collection such as that of Poesio and Vieira (1998).

**3. How to handle functions of referring expressions other than identification?** Target identification is an important function of referring expressions, but it is not the only one. Consider the following example, which Dale and Reiter (1995) discuss to illustrate the limits of their approach:

(5) Sit by *the newly painted table*.

Here, "the newly painted table" allows the addressee to infer that it would be better not to touch the table. To account for examples such as this one, a REG algorithm should be able to take into account different speaker goals (to identify, to warn, etc.) and allow these goals to drive the generation process. These issues were already studied in the plan-based approach to REG of Appelt and Kronfeld (Section 2.1), and more recent work addresses similar problems using new methods. Heeman and Hirst (1995), for example, present a plan-based, computational approach to REG where referring is modelled as goal-directed behaviour. This approach accounts for the combination of different speaker goals, which may be realised in a single referring expression through "information overloading" (Pollack 1991). Context is crucial here: a variant such as "What do you think of the newly painted table?" does *not* trigger the intended "don't touch" inference. In another extension of the plan-based approach to reference, Stone and Webber (1998) use overloading to generate references that only become distinguishing when the rest of the sentence is taken into account. For example, we can say "Take the rabbit from the hat" if there are two rabbits, as long as only one of them is in a hat.

Plan-based approaches to natural language processing are not as popular as they were in the eighties and early nineties, in part because they are difficult to develop and maintain. However, Jordan and Walker (2005) show that a natural language generator can be trained automatically on features inspired by a plan-based model for REG (Jordan 2002). Jordan's "Intentional Influences" model incorporates multiple communicative and task-related problem solving goals, besides the traditional identification goal. Jordan supports her model with data from the Coconut corpus (discussed above) and shows that traditional algorithms such as the IA fail to capture which properties speakers typically select for their references, not only because these algorithms focus on identification, but also because they ignore the interactive setting (see below).

In short, it seems possible to incorporate different goals into a REG algorithm, even without invoking complex planning machinery. However, this calls for a close coupling of REG with the generation of the carrier utterance, containing the generated expression. What impact this has on the architecture of an NLG system, what the relevant goals are, how combinations of different goals influence content selection and linguistic realisation, and how such expressions are best evaluated is still mostly unexplored. Answers might come from studying REG in the context of more complex applications, where the generator may need to refer to objects for different reasons.

**4. How to generate suitable referring expressions in interactive settings?** Ultimately, referring expressions are generated for some addressee, yet most of the algorithms we have discussed are essentially "addressee-blind" (Clark and Bangerter 2004). To be fair, some researchers have paid lip service to the importance of taking the addressee into account (cf. Dale and Reiter's UserKnows function), but it is still largely an open question to what extent the classical approaches to REG can be used in interactions. In fact, there are good reasons to assume that most current REG algorithms cannot directly be applied in an interactive setting. Psycholinguistic studies on reference production, for example, show that human speakers do take the addressee into account when referring (an instance of "audience design" (Clark and Murphy 1983)). Some psycholinguists have argued that referring is an interactive and collaborative process, with speaker and addressee forming a "conceptual pact" on how to refer to some object (Brennan and Clark 1996; Metzing and Brennan 2003; Heeman and Hirst 1995). This also implies that referring is not necessarily a "one shot" affair; rather a speaker may quickly produce a first approximation of a reference to some target, which may be refined following feedback from the addressee.

Others have argued that conversation partners automatically "align" with each other during interaction (Pickering and Garrod 2004). For instance, Branigan et al. (2010) report on a study showing that if a computer uses the word "seat" instead of the more common "bench" in a referring expression, the user is subsequently more likely to use "seat" instead of "bench" as well. This kind of lexical alignment takes place at the level of linguistic realisation, and there is at least one NLG realiser that can mimic this process (Buschmeier, Bergmann, and Kopp 2009). Goudbeek and Krahmer (2010) found that speakers in an interactive setting also align at the level of content selection; they present experimental data showing that human speakers may opt for a "dispreferred" attribute (even when a preferred attribute would be distinguishing) when these were salient in a preceding interaction. The reader may want to consult Arnold (2008) for an overview of studies on reference choice in context, Clark and Bangerter (2004) for a discussion of studies on collaborative references, or Krahmer (2010) for a confrontation of some recent psycholinguistic findings with REG algorithms.

Psycholinguistic findings suggest that traditional REG algorithms which rely on some predefined ranking of attributes cannot straightforwardly be applied in an interactive setting. This is confirmed by the findings of Jordan and Walker (2005) and Gupta and Stent (2005), who studied references in dialogue corpora discussed in Section 5. They found that in these data-sets, traditional algorithms are outperformed by simple strategies which pay attention to the referring expressions produced earlier in the dialogue. More recently, other researchers have started exploring the generation of referring expressions in interactive settings as well. Stoia et al. (2006), for example, presented a system that generates references in situated dialogues, taking into account both dialogue history and spatial visual context, defined in terms of which distractors are in the current field of vision of the speakers and how distant they are from the target. Janarthanam and Lemon (2009) present a method which automatically adapts to the expertise level of the intended addressee (using "the router" when communicating with an expert user, and "the black block with the lights" while interacting with a novice). This line of research fits in well with another, more general, strand of research concentrating on choice optimisation during planning based on user data (Walker et al. 2007; White, Clark, and Moore 2010).

Interactive settings seem to call for sophisticated addressee modelling. However, detailed reasoning about the addressee can be computationally expensive, and some psychologists have argued, based on clever experiments in which speakers and addressees have slightly different information available, that speakers only have limited capabilities for considering the addressee's perspective (Horton and Keysar 1996; Keysar, Lin, and Barr 2003; Lane, Groisman, and Ferreira 2006). Some of the studies mentioned above, however, emphasise a level of cooperation that may not require conscious planning: the balance of work on alignment, for example, suggests that it is predominantly an automatic process which does not take up much computational resource. Recently, Gatt et al. (2011) proposed a new model for interactive REG, consisting of a preference-based search process based on the IA, which selects properties concurrently and in competition with a priming-based process, both contributing properties to a limited capacity working memory buffer. This model offers a new way to think about interactive REG, and the role therein for REG algorithms of the kind discussed in this survey.

**5. What is the impact of visual information?** Throughout this paper we have often discussed references to objects in some shared visual scene, because it offers a useful way to illustrate the workings of an algorithm. Yet only a small handful of REG researchers appear to have taken visual information seriously.

Most real-life scenes contain a multitude of potential referents. Just look around you: every object in your field of vision could be referred to. It is highly unlikely that speakers would take all these objects into account when producing a referring expression. Indeed, there is growing evidence that the visual system and the speech production system are closely intertwined (Meyer et al. (1998), Hanna and Brennan (2007) and Spivey et al. (2001)). Human speakers employ specific strategies when looking at real-world scenes (e.g., Itti and Koch (2000), Wooding et al. (2002)). Wooding and colleagues, for instance, found that certain properties of an image, such as changes in intensity and local contrasts, determine viewing patterns to a large extent. Top-down strategies also play a role: for instance, areas that are currently under discussion are looked at more frequently and for longer periods of time. Yet, little is known about how scene perception influences the human production of referring expressions, and how REG algorithms could mimic this.

When discussing visual scenes, most REG researchers assume that some of the relevant visual information is stored in a database (compare our visual example scene in Figure 1 and its database representation in Table 1). Still, the conversion from one to the other is far from trivial. Clearly, the visual scene is much more informative than the database; how do we decide which visual information we store in the database and which we ignore? And, how do we map visual information to symbolic labels? These are difficult questions, which have received very little attention so far. A partial answer to the first question can be found in the work of John Kelleher and colleagues, who argue that visual and linguistic salience co-determine which aspects of a scene are relevant for the understanding and generation of referring expressions (Kelleher, Costello, and van Genabith 2005; Kelleher and Kruijff 2006). A partial answer to the second question is offered by Deb Roy and colleagues (e.g., Roy and Penland (2002) and Roy (2005)) who present a computational model for automatically grounding attributes based on sensor data, and by Gorniak and Roy (2004) who apply such a model to referring expressions.

One impediment to progress in this area is the lack of relevant human data. Most, if not all, of the dedicated data-sets discussed in Section 5 were collected using artificial visual scenes, either consisting of grids of unrelated objects not forming a coherent scene, or of coherent scenes of unrealistic simplicity. Generally speaking, the situation in psycholinguistics is not much better. Recently, some studies started exploring the effects of more realistic visual scenes on language production. An example is Coco and Keller (2009), who photoshopped a number of (more or less) realistic visual scenes, manipulating the visual clutter and number of actors in each scene. They found that more clutter and more actors resulted in longer delays before language production started, and that these factors influenced the syntactic constructions that were used as well. A similar paradigm could be used to collect a new corpus of human-produced references, with targets being an integral part of a visual scene (rather than being randomly positioned in a grid). When participants are subsequently asked to refer to objects in these scenes, eye tracking can be used to monitor where they are looking before and during the production of particular references. Such data would be instrumental for developing REG algorithms which take visual information seriously.

**6. What Knowledge Representation framework suits REG best?** Recent years have seen a strengthening of the link between REG and knowledge representation frameworks (see Section 4). There is a new emphasis on questions involving (1) the expressive power of the formalism in which domain knowledge is expressed (e.g., does the formalism allow convenient representation of $n$-place predicates or quantification?), (2) the expressive power of the formalism in which ontological information is expressed (e.g., can it express more than just subsumption between concepts?), (3) the amount of support available for logical inference, and (4) the mechanisms available within each framework for controlling the output of the generator.

To illustrate the importance of expressive power and logical inference, consider the type of examples discussed in Poesio and Vieira (1998). What would it take to generate an expression like "the report on the third quarter of 2009"? It would be cumbersome to represent the relation between all entities separately, saying that 1950 has a first quarter, which has a report, and the same for all other years. It would be more elegant and economical to spell out general rules, such as "Every year has a unique first quarter", "Quarter 4 of a given year precedes Quarter 1 of any later year", "The relation "precede" is transitive", and so on. As NLG is starting to be applied in large-scale applications, the ability to capture generalisations of this kind is bound to become increasingly important.

It is remarkable that most REG research has, until now, distanced itself so drastically from other areas of Artificial Intelligence, by limiting itself to atomic facts in the knowledge base. If REG came to be linked with modern knowledge representation formats – as opposed to the simple attribute–value structures exemplified in Table 1 – then atomic formulas are no longer the substance of the knowledge base but merely its seeds. In many cases, resources developed for the semantic web – ontology languages such as OWL, reasoning tools, and even the ontologies themselves – could be re-used in REG. REG could even link up with "real AI", by tapping into models of commonsense knowledge, such as Lenat (1995) or Lieberman et al. (2004). The new possibilities raise interesting scientific and strategic questions. For example, how do *people* generate referring expressions of the kind highlighted by the work of Poesio and colleagues? Is this process best modelled using a knowledge-rich approach using general axioms and deduction, or do other approaches offer a more accurate model? Is it possible that, when REG starts to focus a bit less on identification of the referent, the result might be a different, and possibly less logic-oriented problem? What role could knowledge representation play in these cases? Here, as elsewhere in REG, we see ample space for future research.

## 7. General conclusion and discussion

After preparatory work in the nineteen eighties by Appelt and Kronfeld, and the contributions summarised in Dale and Reiter (1995), the first decade of the new millennium has seen a new surge of interest in referring expression generation. Progress has been made in three related areas which have been discussed extensively in this survey. First, researchers have lifted a number of simplifications present in the work of Dale and Reiter (1995) and others, thereby considerably extending coverage of REG algorithms to include, for instance, relational, plural and vague references (Section 3). Second, proposals have been put forward to recast REG in terms of existing and well-understood computational frameworks, such as labelled directed graphs and Description Logic, with various attractive consequences (Section 4). Last but not least, there has been a shift towards data collection and empirical evaluation; this has made it possible to empirically evaluate REG algorithms, which is starting to give us an improved understanding of the strengths and weaknesses of existing work (Section 5). As a result of these developments, REG is now one of the best developed subfields of NLG.

How should the current state of the art in REG be assessed? The good news is that current REG algorithms can produce natural descriptions, which may even be more helpful than descriptions produced by people (Gatt, Belz, and Kow 2009). However, this is only true when certain simplifying assumptions are made, as in the early REG research typified by Dale and Reiter (1995). When REG leaves this limited "comfort zone", the picture changes drastically. While in recent years the research community has gained a much better understanding of the challenges that face REG in that wider arena, many of these challenges are still waiting to be met (Section 6).

*New complexities.* Recent REG research has revealed various new complexities. Some of these pertain to the nature of the target. *Sets* are difficult to refer to, for example, and algorithms designed to deal with them achieve a lower human-likeness when referring to sets than to individual objects (van Deemter et al. 2011). Recent efforts to let REG algorithms refer to *spatial regions* suggest that in large, realistic domains, precise identification of a target is a goal that can be approximated, but seldom achieved (Turner et al. 2008; Turner, Sripada, and Reiter 2009). Little work has been done so far

on reference to *events*, or to points and intervals in *time* (e.g., "When Harry met Sally", "the moment after the impact"), and references to abstract and other *uncountable entities* (e.g., water, democracy) are beyond the horizon. Where domain knowledge derives from sensor data – with unavoidably uncertain and noisy inputs – this is bound to cause problems not previously addressed by REG. It is in such domains that salience (especially in the non-linguistic sense) becomes a critical issue. When reference takes place in real life – as opposed to a typical psycholinguistics experiment – it is often unclear what their salience depends on. It might be that salience is partly in the eye of the beholder, and that this is one of the causes of the considerable individual variation that exists between different human speakers (Dale and Viethen 2010).

*Human-likeness and evaluation.* In early REG research, including (Dale and Reiter 1995), it was often remarkably unclear what exactly the proposed algorithms aimed to achieve. It was only when evaluation studies were starting to be conducted that researchers had to "show their cards" and say what they regarded as their criterion for success. In most cases, they used a form of human-likeness as their success criterion, by comparing the expressions generated by an algorithm with those in a corpus.

The human-likeness criterion dictates that REG algorithms are to mimic humans "warts and all": if speakers produce unclear descriptions, then so should algorithms. But of course, human-likeness is not the only yardstick that can be used. In NLG systems whose main aim is to be practically useful, for example, it may be more important for referring expressions to be *clear* than to be human-like in all respects. The difference is important because psycholinguists have shown that human speakers have only limited capabilities for taking the addressee into account, frequently producing expressions that cannot be interpreted correctly by an addressee, for example when they are under time pressure (Horton and Keysar 1996). If usefulness, rather than human-likeness is the yardstick for success then a different type of evaluation test needs to be used. Possible tests include, for example, speed and accuracy of task completion (i.e., how often and how fast do readers find the referent?). A variety of hearer-oriented tests is starting to be used in recent REG research (Paraboni, van Deemter, and Masthoff 2007; Khan, van Deemter, and Ritchie 2008), but evaluation of REG algorithms (and of NLG in general) remains difficult, see e.g., Oberlander (1998), Belz (2009) and Gatt and Belz (2010).

Hearer-oriented experiments may also be useful for evaluating referring expressions that are logically complex (cf., Section 4.4). It is one thing for a REG algorithm to use logical quantification to generate a fairly simple description, such as "the woman who feeds four cats", but quite another to generate a highly complex description ("the woman who owns four cats that are chased by between three and six dogs each of which is fed only by men"), which can be generated using the same methods. There are difficult methodological questions to be answered here, about whether the aim of the generator is to model human competence or human performance. And if it is performance that is to be modelled, then this raises the question what types of complexities are exploited by human speakers, and what types of complexities are understandable to human hearers. Such questions can only be answered by new empirical studies.

*Widening the scope of REG algorithms*. Much REG research has concentrated on the main "paradigms" of reference (Searle 1969). Early work on REG treated reference as emphatically part of communication, as we have seen (Section 2.1, First Beginnings). But after the refocussing that went on in the 1990's, many REG algorithms operate as if describing objects were a goal unto itself, instead of a part of communication. Still,

when referring expressions occur in their natural *habitat* – in text or dialogue – then the reference game becomes subtly different, with factors such as salience and adaptation playing important (and partly unknown) roles. In these natural contexts, it is also not always necessary to identify a referent "in one shot": in dialogue, identification of the referent is the joint responsibility of both dialogue partners (e.g., Heeman and Hirst (1995)), and even in monologue, an entire sequence of utterances may guide a hearer towards the referent. In casual conversation, it is even unclear whether exact identification of the referent is a requirement at all, in which case all existing algorithms are wrong-footed. Reference in real life is also characterised by domains that are much larger and complicated than the ones usually studied (at least until they have been narrowed down by means of some salience metric): the set of *people*, for example that we are able to refer to in daily life is almost unlimited, and the properties that we can use to refer to them seem almost unbounded, including not only their physical appearance and location, but their ideas, actions, and so on. Evaluation challenges such as TUNA REG, GREC and GIVE have helped to bring the research community together, focussing on small domains and, predominantly, on simple types of referring expressions. We believe that it is time for evaluation studies to extend their remit and look at the types of complex references that more recent REG research has drawn attention to. Such studies would do well, in our view, to pay considerable attention to the question which referring expressions have the greatest benefit for readers or hearers.

One day, perhaps, all these issues will have been resolved. If there is anything that a survey of the state of the art in REG makes clear it is that, for all the undeniable progress in this growing area of NLG, this holy grail is not within reach yet.

**Acknowledgements** The order of authors was determined by chance; both contributed equally. Emiel Krahmer thanks The Netherlands Organisation for Scientific Research (NWO) for VICI grant "Bridging the Gap between Computational Linguistics and Psycholinguistics: The Case of Referring Expressions" (277-70-007). Kees van Deemter thanks the EPSRC's Platform Grant "Affecting People with Natural Language". We both thank the anonymous reviewers for their constructive comments, and Doug Appelt, Johan van Benthem, Robert Dale, Martijn Goudbeek, Helmut Horacek, Ruud Koolen, Roman Kutlak, Chris Mellish, Margaret Mitchell, Ehud Reiter, Advaith Siddharthan, Matthew Stone, Mariët Theune and especially Albert Gatt, for discussions and/or comments on earlier versions of this text. Thanks to Jette Viethen for her extensive REG bibliography.

**References**

Abbott, Barbara. 2010. *Reference*. Oxford University Press.

Anderson, Anne A., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

Appelt, Douglas. 1985. Planning English referring expressions. *Artificial Intelligence*, 26:1–33.

Appelt, Douglas and Amichai Kronfeld. 1987. A computational model of referring. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 640–647.

Areces, Carlos, Alexander Koller, and Kristina Striegnitz. 2008. Referring expressions as formulas of Description Logic. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 42–49, Salt Fork, Ohio.

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

Baader, Franz, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter

Patel-Schneider. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK.
Baget, Jean-François and Marie-Laure Mugnier. 2002. Extensions of simple conceptual graphs: the complexity of rules and constraints. *Journal of Artificial Intelligence Research*, 16:425 – 465.
Bangalore, Srinivas, Owen Rambow, and Steven Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG)*, pages 1–8, Mitzpe Ramon.
Belz, Anja. 2009. That's nice . . . what can you do with it? (last words). *Computational Linguistics*, 35:111–118.
Belz, Anja and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, OH.
Belz, Anja, Eric Kow, Jette Viethen, and Albert Gatt. 2008. The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 183–191.
Belz, Anja, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The GREC task evaluation challenges. In Emiel Krahmer and Mariët Theune, editors, *Empirical Methods in Natural Language Generation*. Springer Verlag, Berlin, pages 294–327.
Bohnet, Bernd and Robert Dale. 2005. Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1004–1009, Edinburgh.
Brachman, Ronald J. and James G. Schmolze. 1985. An overview of the KL–ONE knowledge representation system. *Cognitive Science*, 9(2):171–216.
Branigan, Holly P., Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42:2355–2368.
Brennan, Susan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.
Buschmeier, Hendrik, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment–capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 82–89.
Callaway, Charles and James Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, Philadelphia.
Chafe, Wallace W. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.
Chantree, Francis, Adam Kilgarriff, Anne de Roeck, and Alistair Willis. 2005. Disambiguating coordinations using word distribution information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.
Clark, Herbert H. and Adrian Bangerter. 2004. Changing ideas about reference. In Ira A. Noveck and Dan Sperber, editors, *Experimental Pragmatics*. Palgrave Macmillan, Basingstoke, pages 25–49.
Clark, Herbert H. and Gregory Murphy. 1983. Audience design in meaning and reference. In Jean Francois Le Ny and Walter Kintsch, editors, *Language and Comprehension*. North Holland, pages 287–299.
Coco, Moreno I. and Frank Keller. 2009. The impact of visual information on reference assignment in sentence production. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci)*, pages 274–279, Amsterdam.
Cohen, Philip R. and Hector J. Levesque. 1985. Speech acts and rationality. In *Proceedings of the 23rd Annual Meeting of the Association of Computational Linguists (ACL)*, pages 49–60, Chicago, Illinois.
Croitoru, Madalina and Kees van Deemter. 2007. A conceptual graph approach to the generation of referring expressions. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2456–2461, Hyderabad, India.
Dale, Robert. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 68–75.
Dale, Robert. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. The MIT Press, Cambridge, Massachusetts.
Dale, Robert and Nicolas Haddock. 1991. Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association of Computational*

*Linguists (EACL)*, pages 161–166, Berlin.
Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.
Dale, Robert and Jette Viethen. 2010. Attribute–centric referring expression generation. In Emiel Krahmer and Mariët Theune, editors, *Empirical Methods in Natural Language Generation*. Springer Verlag, Berlin, pages 163–179.
Denis, Alexandre. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, Trim, Ireland.
DeVault, David, Charles Rich, and Candace L. Sidner. 2004. Natural language generation and discourse context: Computing distractor sets from the focus stack. In *Proceedings of the 17th International Meeting of the Florida Artificial Intelligence Research Society (FLAIRS)*, Miami Beach.
Di Eugenio, Barbara, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 2000. The agreement process: an empirical investigation of human-human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53:1017–1076.
Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302.
Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT)*, pages 138–145.
Engelhardt, Paul E., Karl G.D Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54:554–573.
Gardent, Claire. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 96–103, Philadelphia.
Gardent, Claire and Kristina Striegnitz. 2007. Generating bridging definite descriptions. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning, Volume 3*. Studies in Linguistics and Philosophy, Springer Publishers, pages 369–396.
Garey, Michael R. and David S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP–Completeness*. W.H. Freeman, New York.
Gatt, Albert. 2007. *Generating Coherent References to Multiple Entities*. Unpublished PhD thesis, University of Aberdeen.
Gatt, Albert and Anja Belz. 2010. Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In Emiel Krahmer and Mariët Theune, editors, *Empirical Methods in Natural Language Generation*. Springer Verlag, Berlin, pages 264–293.
Gatt, Albert, Anja Belz, and Eric Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG)*, Salt Fork, Ohio.
Gatt, Albert, Anja Belz, and Eric Kow. 2009. The TUNA–REG challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 174–182, Athens, Greece.
Gatt, Albert, Emiel Krahmer, and Martijn Goudbeek. 2011. Attribute preference and priming in reference production: Experimental evidence and computational modeling. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci)*, Boston, Massachusetts.
Gatt, Albert and Kees van Deemter. 2007. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, 16:423–443.
Gatt, Albert, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG)*, pages 49–56, Schloss Dagstuhl, Germany.
Giuliani, Manuel, Mary Ellen Foster, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2010. Situated reference in a hybrid human-robot interaction system. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 67–76.
Goldberg, Eli, Norbert Driedger, and Richard Kittredge. 1994. Using natural language processing to produce weather forecasts. *IEEE Expert*, 9 (2):45–53.
Gorniak, Peter and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.
Goudbeek, Martijn and Emiel Krahmer. 2010. Preferences versus adaptation during referring expression generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–59, Uppsala, Sweden.
Gratch, Jonathan, Jeff Rickel, Elisabth André, Norman Badler, Justine Cassell, and Eric Petajan.

2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17:54–63.
Grice, Paul. 1975. Logic and conversation. In Peter Cole and Jeffrey L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*. Academic Press, New York, pages 43–58.
Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.
Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.
Guhe, Markus and Ellen Gurman Bard. 2008. Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2404–2409, Austin, TX.
Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and form of referring expressions in discourse. *Language*, 69:247–307.
Gupta, Surabhi and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Copora in Natural Language Generation (UCNLG)*, pages 1–6, Brighton, UK.
Hajičová, Eva. 1993. *Issues of Sentence Structure and Discourse Patterns – Theoretical and Computational Linguistics, Vol. 2*. Charles University, Prague.
Hanna, Joy E. and Susan E. Brennan. 2007. Speaker's eye gaze disambiguates referring expressions early during face–to–face conversation. *Journal of Memory and Language*, 57:596–615.
Heeman, Peter A. and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.
Hendrickx, Iris, Walter Daelemans, Kim Luyckx, Roser Morante, and Vincent Van Asch. 2008. CNTS: Memory–based learning of generating repeated references. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 194–195.
Henschel, Renate, Hua Cheng, and Massimo Poesio. 2000. Pronominalisation revisited. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 306–312, Saarbrücken, Germany.
Hopcroft, John. 1971. An $n$ log($n$) algorithm for minimizing states in a finite automaton. In Zvi Kohave, editor, *Theory of Machines and computations*. Academic Press.
Horacek, Helmut. 1996. A new algorithm for generating referring expressions. In *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI)*, pages 577–581, Budapest, Hungary.
Horacek, Helmut. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 206–213, Madrid.
Horacek, Helmut. 2004. On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG)*, pages 70–79, Brockenhurst, UK.
Horacek, Helmut. 2005. Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)*, pages 58–67, Aberdeen, UK.
Horton, William S. and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59:91–117.
Itti, Laurent and Christof Koch. 2000. A saliency–based search mechanism for overt and covert shifts in visual attention. *Vision Research*, 40:1489–1506.
Jaccard, Paul. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.
Janarthanam, Srinivasan and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 74–81, Athens, Greece.
Jordan, Pamela W. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, University of Pittsburgh.
Jordan, Pamela W. 2002. Contextual influences on attribute selection for repeated descriptions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI Publications, Stanford, CA.
Jordan, Pamela W. and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.
Kelleher, John, Fintan Costello, and Josef van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistics discourse context. *Artificial*

*Intelligence*, 167:62–102.
Kelleher, John and Geert-Jan Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1041–1048, Sydney, Australia.
Kerdiles, Gwen. 2001. *Saying It with Pictures: a Logical Landscape of Conceptual Graphs*. Unpublished PhD thesis, ILLC, Amsterdam.
Keysar, Boaz, Shuhong Lin, and Dale J. Barr. 2003. Limits on theory of mind use in adults. *Cognition*, 89:25–41.
Khan, Imtiaz Hussain, Kees van Deemter, and Graeme Ritchie. 2008. Generation of referring expressions: Managing structural ambiguities. In *Proceedings of the 22th International Conference on Computational Linguistics (COLING)*, Manchester, UK.
Kibble, Rodger and Richard Power. 2004. Optimizing referential coherence in text generation. *Computational Linguistics*, 30:401–416.
Kilgarriff, Adam. 2003. Thesauruses for natural language processing. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLPK)*, pages 5–13.
Koller, Alexander and Matthew Stone. 2007. Sentence generation as a planning problem. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Conference Proceedings (ACL)*, pages 337–343, Prague.
Koller, Alexander, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. The first challenge on generating instructions in virtual environments. In Emiel Krahmer and Mariët Theune, editors, *Empirical Methods in Natural Language Generation*. Springer Verlag, Berlin, pages 328–352.
Koolen, Ruud, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2009. Need I say more? On factors causing referential overspecification. In *Proceedings of the CogSci workshop on the Production of Referring Expressions (PRE-CogSci 2009)*, Amsterdam, The Netherlands.
Kopp, Stefan, Kirsten Bergmann, and Ipke Wachsmuth. 2008. Multimodal communication from multimodal thinking. towards an integrated model of speech and gesture production. *Semantic Computing*, 2:115–136.
Krahmer, Emiel. 2010. What computational linguists can learn from psychologists (and vice versa). *Computational Linguistics*, 36:285–294.
Krahmer, Emiel and Mariët Theune. 2002. Efficient context–sensitive generation of descriptions in context. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*, pages 223 – 264, CSLI Publications, CSLI, Stanford.
Krahmer, Emiel, Mariët Theune, Jette Viethen, and Iris Hendrickx. 2008. Graph: The costs of redundancy in referring expressions. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 227–229, Salt Fork, Ohio.
Krahmer, Emiel, Sebastiaan van Erk, and André Verleg. 2003. Graph–based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.
Kronfeld, Amichai. 1990. *Reference and Computation: An Essay in Applied Philosophy of Language*. Cambridge University Press, Cambridge.
Kumar, Vipin. 1992. Algorithms for constraint satisfaction problems: a survey. *Artificial Intelligence Magazine*, 1:32–44.
Lane, Liane Wardlow, Michelle Groisman, and Victor S. Ferreira. 2006. Don't talk about pink elephants! speakers' control over leaking private information during language production. *Psychological Science*, 17 (4):273–277.
Lenat, Douglas. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communication of the ACM*, 38:33–38.
Lester, James, Jennifer Voerman, Stuart Towns, and Charles Callaway. 1999. Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13:383–414.
Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.
Lieberman, Henry, Hugo Liu, Push Singh, and Barbara Barry. 2004. Beating common sense into interactive applications. *AI Magazine*, pages 63–76.
Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using N–gram co–occurrence statistics. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT–NAACL)*, pages 71–78,

Edmonton, Canada.
Lin, Dekang. 1998. An information–theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 296–304, Madison, Wisconsin.
Lønning, Jan Tore. 1997. Plurals and collectivity. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*. Elsevier, Amsterdam, pages 1009–1054.
Malouf, Robert. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–92.
McCluskey, Edward J. 1965. *Introduction to the Theory of Switching Circuits*. McGraw–Hill, New York.
McCoy, Kathleen and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of ACL Workshop on Discourse and Reference Structure*, pages 63–71, University of Maryland, College Park.
Mellish, Chris, Donia Scott, Lynn Cahill, Daniel Paiva, Roger Evans, and Mike Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12:1–34.
Metzing, Charles A. and Susan E. Brennan. 2003. When conceptual pacts are broken: Partner effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:201–213.
Meyer, Antje S., Astrid M. Sleiderink, and Willem J.M. Levelt. 1998. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66:B25–B33.
Mitchell, Margaret. 2009. Class–based ordering of prenominal modifiers. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 50–57, Athens, Greece.
Nenkova, Ani and Kathleen R. McKeown. 2003. References to named entities: A corpus study. In *Proceedings of the Human Language Technology (HLT) Conference, Companion Volume*, pages 70–73.
Oberlander, Jon. 1998. Do the right thing . . . but expect the unexpected. *Computational Linguistics*, 24:501–507.
O'Donnell, Michael, Hua Cheng, and Janet Hitzeman. 1998. Integrating referring and informing in NP planning. In *Proceedings of the ACL Workshop on The Computational Treatment of Nominals*, pages 46–55, Montreal, Canada.
Olson, David R. 1970. Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77:257–273.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.
Paraboni, Ivandré, Kees van Deemter, and Judith Masthoff. 2007. Generating referring expressions: Making referents easy to identity. *Computational Linguistics*, 33:229–254.
Passonneau, Rebecca. 1996. Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, 39:229–264.
Passonneau, Rebecca. 2006. Measuring agreement on set–valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th Internation al Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.
Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:98–110.
Pickering, Martin and Simon Garrod. 2004. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27:169–226.
Piwek, Paul. 2008. Proximal and distal in language and cognition: Evidence from deictic demonstratives in Dutch. *Journal of Pragmatics*, 40:694–718.
Poesio, Massimo, Rosemary Stevenson, Barbara di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30:309–363.
Poesio, Massimo and Renata Vieira. 1998. A corpus–based investigation of definite description use. *Computational Linguistics*, 24:183–216.
Pollack, Martha. 1991. Overloading intentions for efficient practical reasoning. *Noûs*, 25:513–536.
Portet, Francois, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173:789 – 816.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1980. *A grammar of contemporary English (ninth impression)*. Longman, Burnt Mill, Harlow, Essex.
Read, Ronald C. and Derek G. Corneil. 1977. The graph isomorphism disease. *Journal of Graph*

*Theory*, 1(1):339–363.
Reiter, Ehud. 1990. The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 97–104.
Reiter, Ehud and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 232–238, Nantes, France.
Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
Ren, Yuan, Kees van Deemter, and Jeff Pan. 2010. Charting the potential of Description Logic for the generation of referring expressions. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 115–124.
Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch and Barbara L. Lloyd, editors, *Cognition and Categorization*. Erlbaum, Hillsdale, NJ, pages 27–48.
Roy, Deb. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, 9(8):389–96.
Roy, Deb and Alex Pentland. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146.
Scha, Remko and David Stallard. 1988. Multi-level plurals and distributivity. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 17–24, Buffalo, NY.
Searle, John. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.
Shaw, James and Vasileios Hatzivassiloglou. 1999. Ordering among premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 135–143.
Siddharthan, Advaith and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 407–414, Barcelona, Spain.
Sonnenschein, Susan. 1984. The effect of redundant communication on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research*, 13:147–166.
Sowa, John. 1984. *Conceptual structures: Information Processing in Mind and Machine*. Addison–Wesley.
Spivey, Michael, Melinda Tyler, Kathleen Eberhard, and Michael Tanenhaus. 2001. Linguistically mediated visual search. *Psychological Science*, 12:282–286.
Stoia, Laura, Donna K. Byron, Darla Magdalene Shockley, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Natural Language Generation Conference (INLG)*, pages 81–88.
Stone, Matthew. 2000. On identifying sets. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG)*, pages 116–123, Mitzpe Ramon.
Stone, Matthew, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence*, 19:311–381.
Stone, Matthew and Bonnie Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of the 9th International Workshop on Natural Language Generation (INLG)*, pages 178–187, Niagara–on–the–Lake, Ontario.
Theune, Mariët, Ruud Koolen, Emiel Krahmer, and Sander Wubben. 2011. Does size matter: How much data is required to train a reg algorithm? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, Portland, Oregon.
Turner, Ross, Somayajulu Sripada, and Ehud Reiter. 2009. Generating approximate geographic descriptions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 42–49, Athens, Greece.
Turner, Ross, Somayajulu Sripada, Ehud Reiter, and Ian P. Davy. 2008. Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 16–24.
van Deemter, Kees. 2002. Generating referring expressions: Boolean extensions of the Incremental Algorithm. *Computational Linguistics*, 28(1):37–52.
van Deemter, Kees. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

van Deemter, Kees. 2010. *Not Exactly: In Praise of Vagueness*. Oxford University Press, Oxford, UK.
van Deemter, Kees, Albert Gatt, Ielka van der Sluis, and Richard Power. 2011. Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science, to appear*.
van Deemter, Kees and Emiel Krahmer. 2007. Graphs and Booleans: On the generation of referring expressions. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning, Volume 3*. Studies in Linguistics and Philosophy, Springer Publishers, pages 397–422.
van Deemter, Kees, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG)*, pages 130–132, Sydney, Australia.
van der Sluis, Ielka and Emiel Krahmer. 2007. Generating multimodal referring expressions. *Discourse Processes*, 44(3):145–174.
van der Wege, Mija. 2009. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60:448–463.
van Hentenryck, Pascal. 1989. *Constraint Satisfaction in Logic Programming*. The MIT Press, Cambridge, MA.
van Rijsbergen, C.J. 1979. *Information Retrieval*. Butterworths, London, 2nd edition.
Viethen, Jette and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation (INLG)*, pages 63–70, Sydney, Australia.
Viethen, Jette and Robert Dale. 2007. Evaluation in natural language generation: Lessons from referring expression generation. *Traitement Automatique des Langues*, 48:141 – 160.
Viethen, Jette and Robert Dale. 2008. The use of spatial relations in referring expressions. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 59–67.
Viethen, Jette, Robert Dale, Emiel Krahmer, Mariët Theune, and Pascal Touset. 2008. Controlling redundancy in referring expressions. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.
Viethen, Jette, Simon Zwarts, Robert Dale, and Markus Guhe. 2010. Dialogue reference in a visual domain. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*, Valetta, Malta.
Walker, Marilyn, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research*, 30:413–456.
White, Michael, Robert Clark, and Johanna Moore. 2010. Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, 36:159–201.
Winograd, Terry. 1972. *Understanding Natural Language*. Academic Press, New York.
Wooding, David, Mark Muggelstone, Kevin Purdy, and Alastair Gale. 2002. Eye movements of large populations. *Behavior Research Methods, Instruments and Computers*, 34:509–517.