Generating Referring Expressions: Making Referents Easy to Identify

Ivandré Paraboni*
EACH, University of São Paulo

Kees van Deemter**
Computing Science Department,
University of Aberdeen

Judith Masthoff†
Computing Science Department,
University of Aberdeen

It is often desirable that referring expressions be chosen in such a way that their referents are easy to identify. This article focuses on referring expressions in hierarchically structured domains, exploring the hypothesis that referring expressions can be improved by including logically redundant information in them if this leads to a significant reduction in the amount of search that is needed to identify the referent. Generation algorithms are presented that implement this idea by including logically redundant information into the generated expression, in certain well-circumscribed situations. To test our hypotheses, and to assess the performance of our algorithms, two controlled experiments with human subjects were conducted. The first experiment confirms that human judges have a preference for logically redundant expressions in the cases where our model predicts this to be the case. The second experiment suggests that readers benefit from the kind of logical redundancy that our algorithms produce, as measured in terms of the effort needed to identify the referent of the expression.

1. Introduction

Common sense suggests that speakers and writers who want to get their message across should make their utterances easy to understand. Broadly speaking, this view is confirmed by empirical research (Deutsch 1976; Mangold 1986; Levelt 1989; Sonnenschein 1982, 1984; Clark 1992; Cremers 1996; Arts 2004). The present article will examine its consequences for the generation of referring expressions (GRE). In doing this, we distinguish between two aspects of the "understanding" of a referring expression, which we shall denote by the terms interpretation and resolution. We take **interpretation** to be the process whereby a hearer/reader determines the meaning or logical form of the

* Av.Arlindo Bettio, 1000 - 03828-000, São Paulo, Brazil. E-mail: ivandre@usp.br.
** King's College, Meston building, Aberdeen AB24 3UE, Scotland, UK. E-mail: kvdeemte@csd.abdn.ac.uk.
† King's College, Meston building, Aberdeen AB24 3UE, Scotland, UK. E-mail: jmasthoff@csd.abdn.ac.uk.

Submission received: 17 February 2004; revised submission received: 27 July 2006; accepted for publication: 7 December 2006.

referring expression; we take **resolution** to be the identification of the referent of the expression once its meaning has been determined. It is resolution that will take center stage in our investigation.

Difficulty of resolution and interpretation do not always go hand in hand. Consider sentences (1a)–(1c), uttered somewhere in Brighton but not on Lewes Road. The description in (1a) is longer (and might take more time to read and interpret) than (1b), but the additional material in (1a) makes *resolution* easier once interpretation is successfully completed.

(1a) 968 Lewes Road, Moulsecoomb area

(1b) 968 Lewes Road

(1c) number 968

The first two of these descriptions refer uniquely. As for the third: Lewes Road is a long street. Supposing that other streets in Brighton do not have numbers above 900, then even (1c) is a unique description—but a pretty useless one, because it does not help you to find the house unless your knowledge of Brighton is exceptional. We will explore how a natural-language-generation (NLG) program should make use of logically redundant properties so as to simplify resolution (i.e., the identification of the referent). When we write about identifying or "finding" the referent of a referring expression, we mean this in the sense of determining which object is the intended referent. This conceptual goal may or may not require the hearer to make a physical effort, for example by turning the pages of a book, or more dramatically by walking and waiting for traffic lights.

The fact that referring expressions tend to contain logically redundant information has been observed in many empirical studies. Levelt (1989), for example, mentions the need for redundancy in situations of "degraded communication" (e.g., background noise); and even in normal situations, redundant nondiscriminating information can help the addressee identify the referent (Deutsch 1976; Mangold 1986; Sonnenschein 1982, 1984; Arts 2004). In Levelt's words, psycholinguistic experiments show that

> [l]isteners apparently create a 'gestalt' of the object for which they have to search. It is harder to search for 'something red' than for 'a big red bird', even if the color would be sufficiently discriminating. Information about the *kind* of object to be looked for (e.g., a bird) is especially helpful for constructing such a gestalt. (Levelt 1989, page 131)

Although early GRE algorithms have often followed the Gricean maxim, "be brief" (Grice 1975), by minimizing the number of properties in a generated description, Dale and Reiter (1995) proposed an algorithm that allows certain redundancies, for example, by guaranteeing that each generated description expresses the ontological "type" of the referent, in the form of a noun, a move that addresses Levelt's claim to some extent.[1] In corpus-based studies, it has been shown that logically redundant properties tend to be included when their inclusion fulfils one of a number of pragmatic functions, such as to indicate that a property is of particular importance to the speaker (i.e., it constitutes one of her reasons for being interested in the referent) or to highlight the

1 Dale and Reiter (1995, Section 5) also mention the use of "navigational" (or "attention-directing") information in referring expressions, which they distinguish from "discrimination information," and whose function appears to be to move the attention of the reader/hearer towards an object. The concept is not defined precisely and it is not clear how navigational information should be used in GRE.

speaker's awareness that the referent has the property in question (Jordan 2000, 2002). Implementations of such findings in NLG are not difficult to envisage.

The present article takes this reader-oriented perspective on the redundancy of referring expressions a step further, by asking how a generator can use logically redundant information to reduce the search space within which a reader has to "find" a referent; this will be specifically useful when referents need to be found in situations where the extensions of some of the properties are not known to the reader/hearer in advance (cf., Edmonds [1994] for a related set of problems) and where some effort may be needed to identify the referent. By focusing on the information needs of the hearer/reader, our work, a further development of Paraboni and van Deemter (2002a) that also takes the results of Paraboni, Masthoff, and van Deemter (2006) into account, addresses an issue that lies close to the heart of NLG as a practical enterprise, whose purpose is, after all, to make information accessible to people. These issues originally came to the fore while studying references to parts of documents (Paraboni 2000, 2003; Paraboni and van Deemter 2002a, 2002b) but their relevance extends to many other situations. Our findings will also shed light on the **egocentricity** debate among psycholinguists about the extent to which speakers take hearer's knowledge into account when they speak (Keysar, Lin, and Barr 2003). Throughout the article, we shall focus on issues of Content Determination (as opposed to, for example Lexical Choice), and on the situations in which individuals are first mentioned (as opposed to ones in which linguistic context allows them to be shortened [e.g., Krahmer and Theune 2002; Siddharthan and Copestake 2004]).

2. Ease of Resolution in the Incremental Algorithm

Generation of referring expressions (GRE) is a key task of NLG systems (e.g., Reiter and Dale 2000, Section 5.4). An important aspect of GRE is to find combinations of properties that allow the generator to refer uniquely to an entity, called the **target.** Crucially, GRE algorithms only use properties whose denotations are part of the common knowledge of writer and reader.[2] These algorithms are typically designed in such a way that **generation** is performed quickly (e.g., their worst-case running time tends to be linear [Dale and Reiter 1995; van Deemter 2002]) but the processing effort of the reader is not taken into account. Some algorithms do make a point of generating descriptions that are as brief as possible (Dale 1989), and this can be argued to make interpretation easier. As we have seen, however, in relation to Examples (1a–c), brevity can make resolution difficult.

For concreteness, let us focus on one of the best-known algorithms in this area. The Incremental Algorithm (Dale and Reiter 1995) starts by arranging attributes in a list, after which they are considered one by one, to see if any of their values contributes something to the description, by removing "distractors" (i.e., objects other than the referent); if an attribute (e.g., COLOR) can contribute something, then a suitable value (e.g., RED) for this attribute is selected as part of the description. This is repeated incrementally until the logical conjunction of all selected attribute–value combinations results in a unique identification of the referent. There is no backtracking, and this is what keeps the complexity of the algorithm linear; it is also what causes the algorithm to sometimes express a property *P* even when properties that are added later make *P* logically redundant.

2 A good example of a description failing this requirement occurs in *Get off one stop before I do,* in an exchange between two people who have just met, as a description of where the hearer should get off the bus (Appelt 1985, cited in Dale and Reiter 1995).

Suppose a referring expression identifies its referent uniquely. Then at least two things can stand in the way of finding its referent: the "difficulty" of the individual properties used in the description (i.e., the fact that it may be difficult to ascertain which objects have the property in question [Horacek 2005]), or the size and structure of the search space. To exemplify the first factor, suppose you are queuing up for a concert and want to explain to a friend that a girl further ahead in the queue has his ticket. Color is an attribute that speakers like to use, even if it leads to logical redundancy (Pechmann 1989). This might be done by describing the referent as *the girl in a yellow dress,* or as *the girl with green eyes,* for example. But arguably, the first property contributes more towards your friend's search, because the color of a person's eyes may not leap out at him from afar. In the Incremental Algorithm, the fact that DRESS COLOR is more useful than EYE COLOR could be tackled by letting it precede EYE COLOR in the list of attributes. As a consequence, EYE COLOR would only be considered if the referent cannot be identified uniquely without using a combination of more preferred attributes, including DRESS COLOR. Arguably, this is exactly as it should be, and it shows much of what is good about the Incremental Algorithm. It is not so obvious, however, how the algorithm should deal with the second of the two possible obstacles to resolution: the size and structure of the domain.

3. Problems for Resolution

In this section we shall introduce a class of domains (Section 3.1) and a class of problems for resolution that can arise when objects in these domains are identified using a distinguishing description (Section 3.2). Section 4 will relate these problems to a simple model of the resolution process and propose a remedy, which consists of generating logically redundant descriptions (in two different ways). Sections 5 and 6 provide examples of putting our ideas to the test: first, by investigating what kind of description is preferred by subjects who are given the choice (Section 5); then, more elaborately, by investigating the effect of redundant descriptions on readers (Section 6).

3.1 Hierarchical Domains

Existing work on GRE tends to focus on fairly simple domains, dominated by one-place properties. When relations (i.e., two-place properties) are taken into account at all (e.g., Dale and Haddock 1991; Krahmer and Theune 2002), the motivating examples are kept so small that it is reasonable to assume that speaker and hearer know all the relevant facts in advance. Consequently, search is not much of an issue (i.e., resolution is easy): The hearer can identify the referent by simply intersecting the denotations of the properties in the description, for example, intersecting the set of girls with the set of individuals who wear a yellow dress (both in the domain). Although such simplifications permit the study of many aspects of reference, other aspects come to the fore when larger, and subtly structured, domains are considered.

Interesting questions arise, for example, when a large domain is **hierarchically ordered**. For the purpose of this article, we consider a domain to be hierarchically ordered if its inhabitants can be structured like a tree in which everything that belongs to a given node n belongs to at most one of n's children, and everything that belongs to one of n's children belongs to n. Examples include countries divided into provinces, which, in turn, may be divided into regions, and so on; years into months, then into weeks, and then into days; documents into chapters, then sections, then subsections;

buildings into floors, then rooms. Clearly, hierarchies are among our favorite ways of structuring the world.[3]

A crucial question, in all such cases, is what knowledge is shared between speaker and hearer at utterance time. Later on (most explicitly in Section 6), we shall focus on more realistic situations but, to get the idea, it will be useful to think about the extreme case where, before the start of resolution (i.e., before consulting the "knowledge in the world," as opposed to the hearer's "knowledge in the head" [Norman 1988]), the hearer knows nothing about the domain. When the utterance is made, the hearer's blindfold is removed, so to speak, and resolution can start. No similar assumption about the speaker is made: We assume that the speaker knows everything about the domain, and that he knows that the hearer can achieve the same knowledge. Many of our examples will be drawn from a simple model of a University campus, structured into buildings and rooms; the intended referent will often be a library located in one of the rooms. The location of the library is not known to the hearer, but it is known to the speaker.

Each domain entity r will be associated with a TYPE (e.g., the type 'room'), and with some additional attributes such as its ROOM NUMBER or NAME, and we will assume that it is always possible to distinguish r from its siblings in the tree structure by using one or more of these properties. (For example, ROOM NUMBER = 120 identifies a room uniquely within a given building; BUILDINGNAME = Watts identifies a building within the university.) This is a useful assumption, because without it, the existence of a distinguishing description cannot be guaranteed.

The kinds of referring expression that we are interested in (see Section 5 for motivation) take the form of a list

$$L = \langle (x_1, P_1), (x_2, P_2) \ldots (x_n, P_n) \rangle$$

where $x_1 = r$ is the referent of the referring expression and, for every $j > 1$, x_j is an ancestor (not necessarily the parent) of x_{j-1} in the domain D. For every j, P_j is a set of properties that jointly identify x_j within x_{j+1} or, if $j = n$, within the whole domain. The reference *the library in room 120 of Cockcroft building*, for example, is modeled as

$$L = \langle (r, \{type = library\}), (x_2, \{type = room, roomnumber = 120\}), (x_3, \{type = building, buildingname = Cockcroft\}) \rangle$$

3.2 Obstacles for Resolution

We have argued that generating a uniquely referring expression is not always enough, because such an expression can leave the hearer with an unnecessarily large search space. But the issue is an even starker one, especially—as we shall soon see—when it is taken into account that references in hierarchically structured domains can make use of the position of the speaker and hearer in the domain. (For simplicity, we assume that these two locations coincide.)

Let us start with some informal observations, to be corroborated in Section 4. Suppose a hierarchically ordered domain D contains *only one* entity whose TYPE is LIBRARY.

3 If everything that belongs to a given node n belongs to exactly one of n's children, then nodes can be thought of as being partitioned by its children. Note that this is not always the case. Not everything on a given floor of a building, for example, has to be in a room (the corridors are not).

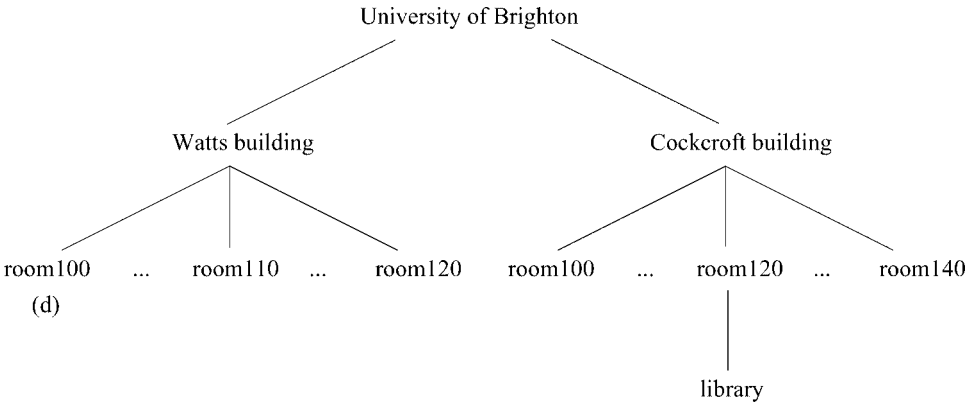

Figure 1
A hierarchically structured domain; *d* is where the reference is uttered.

Consider the following noun phrases, uttered in the position marked by *d* in Figure 1. (The first three have the same intended referent.)

(2a) the library, in room 120 in the Cockcroft building

(2b) the library, in room 120

(2c) the library

(2d) room 140

Utterances like Examples (2a) and (2b) make use of the hierarchical structure of the domain.[4] We focus on the search for x_n (i.e., the highest hierarchical level referred to in the description) because, under the assumptions that were just made (in particular the fact that x_j be identified uniquely in x_{j+1} by the properties P_j), this is the only place where problems can be expected (because no parent node is available).

Even though each of Examples (2a)–(2d) succeeds in characterizing their intended referent uniquely, some of these descriptions can be problematic for the hearer. One type of problem occurs in Example (2d). The expression is logically sufficient (i.e., there is only one room labeled *140* in the entire university). But, intuitively speaking, the expression creates an expectation that the referent may be found nearby, within the Watts building, whereas, in fact, a match can only be found in another building. In a case like this, we will speak of **Lack of Orientation** (*LO*). Even more confusion might occur if another library was added to our example, for instance in Watts 110, whereas the intended referent was the other library (i.e., in room 120 Cockcroft). In this case, Example (2c) would misfire, of course. The expression (2b), however, would succeed, by *mutually* using two parts of the description (*the library* and *room 120*) to identify another: There are two libraries, and two rooms numbered 120, but there is only one pair (a, b) such that a is a library and b is a room numbered 120, with a located in b. Such cases of

4 Recall that we focus on Content Determination, bypassing issues to do with lexical choice, linguistic realization, and so on. For example, we shall not worry whether it is better to say (i) *the library, in room 120*, (ii) *the library in room 120* (without a comma), or (iii) *the library (room 120)*. The difference is not trivial, because (ii), for example, might be viewed as having the unwanted implicature that there is more than one library in the Watts building (Robert Dale, personal communication, August 2005.)

mutual identification[5] are unproblematic in small, transparent, domains where search is not an issue, but in large hierarchical domains, they are awkward (see the Conclusion). For, like Example (2d), (2b) would force a reader to search through an unnecessarily large part of the domain; worse even, the search "path" that the reader is likely to follow leads via an obstacle (namely, room 120 Watts) that matches a part of the description, although not being the intended referent of the relevant part of the description (i.e., room 120 Cockcroft). Confusion could easily result. For even if the reader eventually finds the library, she has no simple way of knowing whether it is the right one. (Perhaps a library in Watts 120 has been overlooked.) In cases like this, we speak of a **Dead End** (*DE*).

Suppose the domain D is represented as a finite tree whose nodes have attributes associated with them, one of which is the TYPE attribute. As before, we shall assume that the attributes and values suffice to identify every node within its parent node. Before defining *LO* and *DE* more precisely, we describe the related notions of SCOPE and SCOPEGROUP, and the notion of a **search path**. We write $x \in D$ to say that x is a node in the tree D; if A is an attribute applicable to x then $A(x)$ denotes the value of A for x.

> **Scope**: Suppose $x \in D$, and $A_1, \ldots, A_n$ are attributes associated with x. Then SCOPE$(x, \{A_1, \ldots A_n\})$ is the largest subtree S of D such that $x \in S$ while, for every $y, z \in S$, the conjunction $A_1(y) = A_1(z)$ & $\ldots$ & $A_n(y) = A_n(z)$ implies $y = z$.

SCOPE$(x, \{A_1, \ldots A_n\})$ is the largest subtree of D in which the values for the attributes $A_1, \ldots A_n$ jointly succeed in pinning down the referent. In practice, we shall usually focus on situations where $n = 1$, in which case we shall write SCOPE(x, A_1), omitting the brackets. In our University domain, let x be room 140 of Cockcroft, then SCOPE(x, ROOM NUMBER) is the subtree rooted in Cockcroft, because within Cockcroft, all room numbers are unique, whereas at the level of the entire university (the next level up), this is not the case (even though the room number 140 itself happens to be unique at that level).

The notion of SCOPE gives rise to the notion of SCOPEGROUP in a straightforward way. Assuming, once again, that $x \in D$, and letting U stand for a set of attributes associated with x, we define:

$$\text{SCOPEGROUP}(x, U) = \{y \in D \mid y \in \text{SCOPE}(x, U) \;\&\; \text{TYPE}(x) = \text{TYPE}(y)\}$$

Thus, SCOPEGROUP $(x, \{A_1, \ldots A_n\})$ is the set of those elements of SCOPE$(x, \{A_1, \ldots A_n\})$ that are of the same TYPE as x. Again, we shall focus on cases where $n = 1$, and omit brackets. Thus, in the example domain, SCOPEGROUP(x,ROOM NUMBER), where x is any room in Cockcroft, is the set of all the rooms in Cockcroft. TYPE is kept constant in the definition of SCOPEGROUP because it tends to be the only non-structural attribute that is used to identify domain objects (i.e., the only attribute that is not intended for designating a node of the domain tree).[6] Non-structural attributes will be assumed to be unproblematic, operating like a filter on the set of possible referents. For example, a reader of the description *the library in room 110* will only be looking for libraries (although they might be looking for them in the wrong building).

5 A well-known example is the description *the bowl on the table*, in a domain that contains several tables and several bowls, but only one bowl on a table (Dale and Haddock 1991).

6 For example, we have seldom found descriptions like "the section *containing tables*," "the *italicized* section" in the PILs corpus (ABPI 1997).

We are now in a position to define *DE* and *LO* more precisely, relative to a **search path.** A search path is a series of steps in the search for a referent, representing visits to nodes in the domain tree D. The path will be modeled by an ordered list of visited nodes: $O = \langle n_1, n_2, \ldots n_m \rangle$. The node n_1 is visited first, then n_2, and so on, until either the referent is found (success) or the reader gives up (failure). As before, let $L = \langle (x_1, P_1), (x_2, P_2) \ldots (x_n, P_n) \rangle$ model the semantic structure of the description, in which x_n is the entity of highest hierarchical level referred to in L. Furthermore, let $A_1, \ldots, A_j$ be the set of all attributes in P_n. Then we predict problems for resolution to occur if some y occurs prior to x_n in O, for which $\text{TYPE}(x_n) = \text{TYPE}(y)$ and $y \notin \text{SCOPEGROUP}(x_n, A_1, \ldots A_j)$. Calling such y an **obstacle,** there are two types of obstacle: the obstacles for which all the properties in P_n are true (these are perhaps the worst kind, because they can be mistaken for the intended referent), and the ones for which this is not the case. If only obstacles of the latter kind arise then we will speak of Lack of Orientation (*LO*). If there is at least one obstacle of the former, more serious kind, we will speak of Dead End (*DE*). For example, in the case of the *DE* Example (2b) (*the library in room 120*), the description itself can be modeled as the list $L = \langle (r, \{type = library\}), (x_2, \{type = room, roomnumber = 120\}) \rangle$, where P_n is the property ROOM NUMBER = 120 and x_n is the room where the library is. Suppose that the search path for x_n corresponds to the following sequence (because referents are always found in leaf nodes, other nodes appear in brackets):

$O = \langle$*Watts*100, (*Watts*,)*Watts*110, (*Watts*,)*Watts*120, (*Watts*,)(*University*,)
(*Cockcroft*,)*Cockcroft*100, (*Cockcroft*,)*Cockcroft*120$\rangle$

Part of this sequence is the obstacle $y =$ Watts 120, which is of the same TYPE as x_n (i.e., both are rooms), and which does not belong to SCOPEGROUP(x_n, ROOM NUMBER) (i.e., it does not belong to the Cockcroft building).

Because the property P_n (ROOM NUMBER = 120) is true of y, this constitutes a case of *DE*. If the room Watts 120 is removed from the domain, there no longer exists an obstacle of the most serious kind (because there is only one room whose room number is 120), but rooms 100 and 110 in the Watts building *are* obstacles of the less serious kind, making this an example of *LO*.

It seems likely that *DE*s and *LO* can disrupt search in sufficiently large or complex domain structures. In principle, *DE* and *LO* could result even in the most unlikely regions of the domain. Suppose *the cup on the table* is uttered in a room d, which contains the intended referent. Now suppose (rather perversely perhaps) the hearer started searching in *another* room, say the kitchen, before looking at the nearest table (in d). If the kitchen happens to contain a table as well, and this table does not support any cups, *DE* would result. Search, however, seems unlikely to proceed in this way. To make testable predictions, we will make some assumptions concerning the way in which referring expressions are resolved by hearers. To explain what these assumptions are, let us return to the examples in Section 3, repeated here for convenience.

(2a) the library in room 120 in the Cockcroft building

(2b) the library in room 120

(2c) the library

(2d) room 140

We assume that these sentences are uttered in the University, say at the location d, and that d determines the starting point of the search for a referent. Henceforth the starting point s will be assumed to be the parent node of d. The intuition behind this assumption is simple: When searching, start looking *nearby*.

It will often be useful to assume that resolution adheres to a principle that we will call **Ancestral Search**. In formulating this principle, we will use d' as a name for the referring expression (which, as we know, takes place at location d); we will use Ref(d') as short for *the intended referent of* d'.

> **Ancestral Search:** First, search for Ref(d') in the subtree dominated by the starting point s. If Ref(d') is not found there then search for Ref(d') in the subtree dominated by the parent of s, which is called s'. If Ref(d') is not found there then move up to the parent s'' of s',..., and so forth, until the root is reached. [If, at this point, Ref(d') is still not found, search fails.]

Ancestral Search (*AS*) says that the hearer of a referring expression searches exhaustively through the current search space (e.g., the building in which the expression is uttered, or the current document section containing the expression) before inspecting a larger subtree. *AS* does not say how the search *within each subtree* (i.e., the one dominated by s or s') is carried out. We do not claim that readers always adhere exactly to *AS*, especially not when they are confronted with unusual situations (as we shall see in our second experiment). Rather, *AS* can be seen as an "ideal model," much like a straight line could be seen as an ideal model of how a pedestrian walks from one point to another. We shall see later that *AS* makes surprisingly accurate predictions in terms of what references are found difficult by readers.

4. Generation Algorithms

What kinds of expression would existing GRE algorithms produce in the situations of interest? Because hierarchies involve relations, the first algorithm that comes to mind is the one proposed by Dale and Haddock (1991). Essentially, this algorithm combines one- and two-place predicates, until a combination is found that pins down the target referent. A standard example involves a domain containing two tables and two bowls, although only one of the two tables has a bowl on it. In this situation, the combination $\{bowl(x), on(x,y), table(y)\}$ identifies x (and y as well), because only one value of x can verify the three predicates, and this justifies the description *the bowl on the table*. Now consider Figure 2, with one additional library in room 110 of the Watts building. Here the combination $\{library(x), in(x,y), room(y), roomnumber(y) = 120\}$ identifies x (and y too), because no other library is located in a room with room number 120 (and no other room numbered 120 contains a library). Thus, the standard approach to relational descriptions allows precisely the kinds of situation that we have described as *DE*. Henceforth, we shall describe this as the **Minimal Description** (*MD*) approach to reference because, in the situations of interest, it uses the minimum number of properties by which the referent can be distinguished.

Another option would be to treat a relation like "being in room 120" as a one-place property of the library, and to use the Incremental Algorithm (Dale and Reiter 1995) to generate the descriptions in question. This, however, would not produce results that are interestingly different from *MD*. Suppose, for example, that the TYPE attribute is most preferred (i.e., considered first by the algorithm), with values such as 'library', 'room', and so on. Suppose, furthermore, that the attribute ROOM NUMBER is preferred over

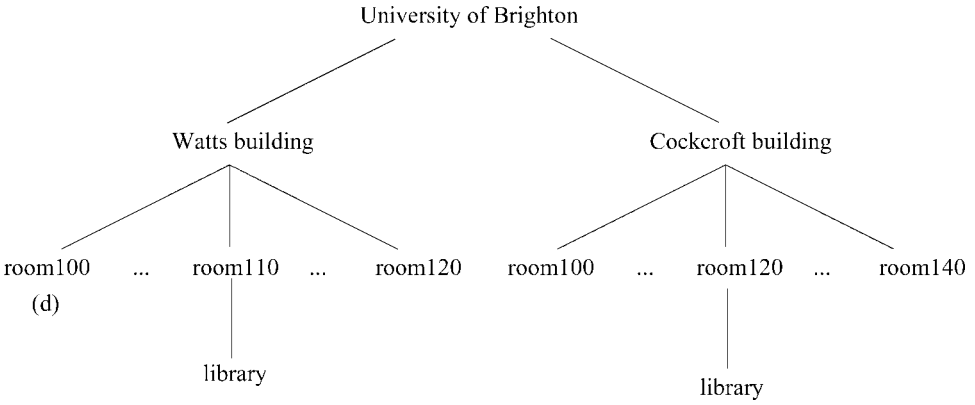

Figure 2
A university campus with two libraries in different buildings.

the attribute BUILDING NAME and, crucially, that a property such as ROOM NUMBER = x is interpreted as true of all those objects in the university (regardless in which building) that are located in something whose room number is x. Then the Incremental Algorithm starts selecting TYPE = library, followed by ROOM NUMBER = 120, at which stage a distinguishing description is reached. In other words, the same description would be generated by this algorithm as by Dale and Haddock (1991) and, once, again, the infamous *LO* and *DE* would occur. Choosing a preference order in which building names are preferred over room numbers would produce *the library in Cockcroft.* Although this description seems defensible in this case, it is easy to see that this preference order would produce excessively lengthy descriptions in other situations. No single preference order produces acceptable results in all cases.

We will now sketch two GRE algorithms, both of which are guaranteed to prevent *DE* and *LO* if *AS* holds. (These algorithms will be investigated empirically in Sections 5 and 6.) They operate by reducing the reader's search space, including logically redundant information into the descriptions that they generate. These algorithms, called **Full Inclusion** (*FI*) and **Scope-Limited** (*SL*), are not the only ways in which resolution may be aided, but we will see that they represent two natural options. Both take as input a hierarchical domain D, a location d where the referring expression will materialize, and an intended referent r. The output is a list of properties L to be turned into an English description by a language realization program.

The first algorithm, *FI*, represents a straightforward way of reducing the length of search paths, without particular attention to *LO* or *DE*. It lines up properties that identify the referent uniquely within its parent node, then moves up to identify this parent node within its parent node, and so on until reaching a subtree that includes the starting point d.[7] *FI* may be likened to existing treatments of salience. In Krahmer and Theune's (2002) approach to GRE, for example, distractors that have lower salience than the intended referent do not have to be removed. We apply this idea to hierarchical domains using the assumption that from the point d where the utterance was made all nodes within d's parent node are as salient as d itself, while more "distant" nodes are gradually less salient. As in Krahmer and Theune, salience sometimes allows for shorter

7 "Includes" is taken to be reflexive: a includes b iff a is an ancestor of b or $a = b$.

descriptions, as when *room 110* replaces *room 110 in Watts* when said in Watts building (but outside room 110).

Full Inclusion(r):

$L := \langle\rangle$ { Initialize L as the empty list }

FI.Identify(r)

The function FI.Identify is defined recursively: (For simplicity, L does not contain the individual referents $x_1,\ldots,x_j$, but only their properties.)

FI.Identify(X):

$L := L + P$, where P identifies X uniquely within *Parent*(X)

$X :=$ *Parent*(X)

IF X includes d THEN STOP ELSE FI.Identify(X)

Applied to our earlier example of a reference to room 120, *FI* first builds up the list $L = \langle(type = room, roomnumber = 120)\rangle$, then expands it to $L = \langle(type = room, roomnumber = 120), (buildingname = Cockcroft)\rangle$. Now that *Parent*($X$) includes d , r has been identified uniquely within D and we reach STOP. L might be realized as *room 120 in Cockcroft,* for example.

FI gives maximal weight to ease of resolution. But something has to give, and that is brevity: By conveying logically redundant information, descriptions are lengthened, and this can have drawbacks, most evidently when there are limitations of space or time. The second algorithm, called *SL*, constitutes a compromise between brevity and ease of resolution. *SL* prevents *DE* and *LO* but opts for brevity when *DE* and *LO* do not occur. Put differently, *SL* favors ease of *resolution* when there is a risk of *DE* or *LO*, but ease of *interpretation* when there is no such risk. This is done by making use of the notion of SCOPE, which was used in the definition of *DE* and *LO*. It may be recalled that a description (x, P) in which P conveys attributes $A_1,\ldots A_j$ leads to *DE* or *LO* when its hearer comes across a node of the same type as x that is not a member of SCOPEGROUP($x, \{A_1,\ldots A_j\}$). It follows that when the hearer is searching within SCOPE($x, \{A_1,\ldots A_j\}$), the description (x, P), even if minimally distinguishing, cannot lead to *DE* or *LO*. Consequently, (x, P) can be uttered in any position d within the subtree denoted by SCOPE($x, \{A_1,\ldots A_j\}$) with no risk of leading to *DE* or *LO* situations. In other words, if SCOPE($x, \{A_1,\ldots A_n\}$) contains d, and if $A_1,\ldots A_n$ are the attributes conveyed in a description (x, P), then this description does not lead to *DE* or *LO*. This allows *SL* to use logically redundant properties more sparingly:

Scope-Limited(r):

$L := \langle\rangle$ { Initialize L as the empty list }

SL.Identify(r)

SL.Identify(X):

$L := L + P$, where P identifies X uniquely within *Parent*(X)

$X :=$ *Root*(*Scope*(X, $\{A_1, \ldots, A_j\}$)), where $A_1, \ldots, A_j$ are the attributes associated with P

IF X includes d THEN STOP ELSE SL.Identify(X)

Whereas *FI* only terminates the generation of the description when a node that includes d is reached, *SL* concludes potentially much earlier, when an attribute (or a combination of attributes) is used that is guaranteed to identify all objects of the relevant type uniquely throughout a tree that includes d. By taking scope into account, *SL* avoids the inclusion of any hierarchical levels not strictly required for preventing *DE* and *LO*.

Consider a description uttered in the position $d =$ room 100 of Watts, with $r =$ room 140 (in Cockcroft) as the intended referent. Existing GRE approaches such as Dale and Reiter (1995) would tend to produce a minimally distinguishing description such as *room 140,* causing *LO*. *SL*, by contrast, would produce the description *room 140 in Cockcroft*,[8] which in this case is the same description produced by *FI*. The *difference* between *FI* and *SL* becomes evident when we consider a case in which the minimally distinguishing description does not lead to *DE*/*LO*—that is, when *AS* predicts that the reader will meet no *DE* or *LO* obstacles. For example, let's return to the situation depicted in Figure 1, from Section 3.1, where there is only one library in the whole university. A reference to $r =$ library would be realized by *FI* as *the library in room 120 in Cockcroft.* By using *SL*, however, the same description would be realized simply as *the library,* because the SCOPE of the attribute TYPE is the whole domain tree [more precisely, SCOPE(LIBRARY,ROOM NUMBER)$= D$] because there is only one entity of TYPE 'library' in the domain and hence no other properties are added. Note that the addition of a second library in the Watts building would reduce SCOPE(r,TYPE) to the subtree rooted in the 'building' node (i.e., each library would be defined by the building to which it belongs). The behavior of the *SL* algorithm would change accordingly, producing *the library in Cockcroft.* Similarly, had we instead included the second library under another room of Cockcroft, the SCOPE would have been reduced even further, causing *SL* to describe r as *the library in room 120 of Cockcroft,* just like the *FI* algorithm.

5. First Experiment: Measuring Reader's Preferences

In this section we start putting the intuition that *LO* and *DE* are better avoided to the test. We report on a small experiment with human subjects, which involved a *document* structured in sections and subsections as an example of a hierarchically ordered domain. We chose this domain because, unlike most other domains, it allows us to show subjects the domain itself (i.e., a real document), rather than, for example, a pictorial representation of it. More specifically, we investigated the choice of so-called **document-deictic** references, such as *the picture in part x of section y* (Paraboni 2003), to check whether they avoid potential *DE* and *LO* situations by adding logically redundant properties (favoring ease of resolution) and, conversely, whether they choose shorter descriptions when there is no such risk (favoring ease of interpretation).

8 The reason is that *Root*(*Scope*(r,ROOM NUMBER)) = Cockcroft, which does not include d. This causes the algorithm to have to identify the Cockcroft building before the algorithm stops.

5.1 Experiment Design

Subjects. 15 academics with considerable practice in the authoring of papers on computational linguistics.

Procedure. A within-subjects design was used. All subjects were shown a printed document containing 18 incomplete statements. Subjects were asked to put themselves in the shoes of the author and to choose the description that they found more suitable for each situation:

> Suppose you and a colleague are currently collaborating on this document. Fortunately he/she did almost all the work for you, and now all that you have to do is complete certain parts of the existing text [. . .]

Subjects completed the statements by choosing one of two alternatives provided: one "minimally distinguishing" description and the other conveying logical redundancy (corresponding to the output of the *FI* or *SL* algorithms). Both alternatives are unambiguous references to the same object. Figure 3 shows a number of descriptions of this kind (whose intended referents are elsewhere in the document) and objects (referred to by descriptions elsewhere). Statement 11 gives a choice between a logically redundant description as generated by *FI* or *SL* (*Part C of Section 2*) and its minimally distinguishing alternative *Part C*. Both alternatives are unambiguous because there is only one part labeled as "C" in the document, but the shorter one may potentially lead to *LO* because the current document section does not contain a part labeled as "C". Similarly, Statement 12 gives a choice between a minimally distinguishing description as generated by *MD* or *SL*, and a logically redundant alternative as generated by *FI*, but in this case none of the alternatives can lead to *DE* or *LO* because there is only one "table 2" in the entire document. The presentational order of alternatives

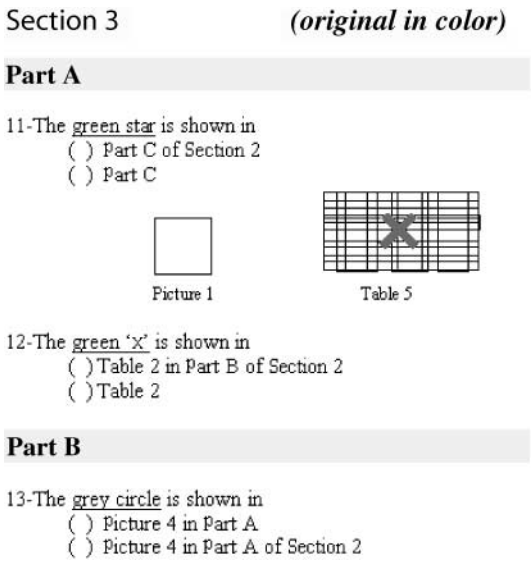

Figure 3
Fragment of the document used in the experiment.

Table 1
Situations of reference for Experiment 1.

Sit.	Type	Reader Loc.	Referent Loc.	MD	Redundant
2	DE	Part A Sec 1	Part B Sec 3	*Pic 2 in Part B*	*Pic 2 in Part B Sec 3*
9	DE	Part C Sec 2	Part B Sec 3	*Pic 3 in Part B*	*Pic 3 in Part B Sec 3*
13	DE	Part B Sec 3	Part A Sec 2	*Pic 4 in Part A*	*Pic 4 in Part A Sec 2*
15	DE	Part B Sec 3	Part A Sec 2	*Pic 3 in Part A*	*Pic 3 in Part A Sec 2*
5	LO	Part B Sec 1	Part A Sec 2	*Pic 5*	*Pic 5 in Part A Sec 2*
7	LO	Part B Sec 1	Part C Sec 2	*Part C*	*Part C Sec 2*
11	LO	Part A Sec 3	Part C Sec 2	*Part C*	*Part C Sec 2*
16	LO	Part B Sec 3	Part A Sec 2	*Pic 6*	*Pic 6 in Part A Sec 2*
4	NONE	Part A Sec 1	Part B Sec 3	*Table 6*	*Table 6 in Part B Sec 3*
10	NONE	Part C Sec 2	Part A Sec 3	*Table 5*	*Table 5 in Part A Sec 3*
12	NONE	Part A Sec 3	Part B Sec 2	*Table 2*	*Table 2 in Part B Sec 2*
18	NONE	Part B Sec 3	Part A Sec 1	*Table 1*	*Table 1 in Part A Sec 1*

(i.e., short versus redundant descriptions) was evenly distributed, to control for order effects.

Research questions. We were interested in seeing whether readers prefer longer (i.e., logically redundant) descriptions when there is a risk of *DE* or *LO* and, conversely, whether they prefer minimally distinguishing descriptions when there is no such risk. Table 1 shows the type of situation (potential *DE*, *LO*, and non-problematic), the reader and referent location, and the descriptions used. To break the monotony of the task and to disguise the purpose of the experiment, another six situations were used that were not relevant to the experiment. Half of the situations, in each of the types, involved backward references, the other half involved forward references. Pictures were enumerated per part so that we could compare short and long versions of potentially problematic descriptions (e.g., *Picture 5* in which the intended referent is not in the current document part, which may or may not contain other pictures). Within the *LO* situations, two of the four statements involved references to pictures, whereas the other two involved references to sections. This was done in order to test whether the type of referent had any influence on the choices made by the subjects. All the questions related to potential *DE* situations involved references to pictures, because using *DE* references to sections would have led to highly artificial structures.

> **Hypothesis 1.1:** In a problematic *DE* situation, descriptions generated by *FI* or *SL* are preferred over minimally distinguishing (*MD*) descriptions.

We will use the *DE* situations in Table 1 to test this hypothesis, investigating how often subjects prefer *FI*/*SL* descriptions to *MD* ones.

> **Hypothesis 1.2:** In a problematic *LO* situation, descriptions generated by *FI* or *SL* are preferred over minimally distinguishing (*MD*) descriptions.

We will use the *LO* situations in Table 1 to test this hypothesis, investigating how often subjects prefer *FI*/*SL* descriptions to *MD* ones. Note that in problematic situations, *SL* generates the same descriptions as *FI*.

We also wanted to investigate whether subjects would prefer descriptions generated by *FI* or *SL* in non-problematic situations (i.e., those not involving potential *DE* or *LO*). We did not use pictures as we did in the problematic cases because in these cases both *FI* and *SL* would produce the same descriptions (e.g., *Picture 5*).[9] In order to compare these algorithms in non-problematic situations we used tables enumerated throughout the document, in which case descriptions produced by *SL* are short (e.g., *Table 5*) and descriptions produced by *FI* are longer (e.g., *Table 5 in Part C of Section 2*).

Hypothesis 1.3: In a non-problematic situation (i.e., a situation not involving *DE* or *LO*), *SL* or *MD* descriptions are preferred over those generated by *FI*.

We will use the non-problematic situations in Table 1 to test this hypothesis, investigating how often subjects prefer *FI* descriptions to *MD*/*SL* ones. Note that in these non-problematic situations, *SL* generates the same descriptions as *MD*.

Hypotheses 1.1 and 1.2 investigate whether ease of resolution (as in logically redundant descriptions generated by *FI* or *SL*) is favored over ease of interpretation (as in minimally distinguishing descriptions) when the description may lead to *DE* or *LO*. Hypothesis 1.3 investigates whether ease of interpretation (as in *MD* or *SL* descriptions) is favored over ease of resolution (as in descriptions generated by *FI*) when the former does not lead to *DE* or *LO* situations.

*Materials. DE*s and *LO* can only occur in fairly complex domains. Instead of trying to find a large number of such documents, we made use of a specially designed schematic document. The document was presented in a printed version (3 pages long), divided into sections (1–3) and subsections ("A" and "B"); Section 2 contained also a subsection labelled "C".[10]

References to pictures can be realized in many different ways. For example, the referent can be called *Picture* or *Figure* or just *Fig.;* the reference can be constructed from the bottom up (*Picture 3 in Section 4*) or from the top downwards (*Section 4, Picture 3*); punctuation varies as well, as does the use of capitals. In our experiments, we have made one fairly arbitrary choice from among all these possibilities, motivated by the types of reference that we observed most frequently in an informal study of a collection of patient information leaflets from the PILs corpus (ABPI 1997): We always used the word *Picture,* we constructed the references bottom up (going up one level at a time), and never used commas or semicolons. Thus, for example, we asked subjects to compare *Picture 3 in Part B of Section 3* with *Picture 3 in Part B.*[11] Even though it is *possible* that a different realization choice would produce different experimental outcomes, this does not seem likely.

Every description *d* and its referent *r* were always on different pages. Had *d* and *r* occurred on the same page then physical proximity might have obscured navigational

9 In the second experiment reported in Section 6 this was no longer an issue as we focus on ease of resolution only, that is, it did not compare *FI* with *SL*.

10 See Paraboni (2003), appendix 1, for the actual document.

11 To get a feeling for the frequency of the expressions involved, one might enter "picture OR figure OR fig 1…9 in part OR section 1…9" into Google, using *Advanced Search*. In July 2006, this produced as many as 77,000 hits, the great majority of which are of the intended kind. (Because Advanced Search disregards punctuation and capitalization, this includes a very small percentage of false positives, for example of the form "Figure *x*. In section *y* …".) The materials of our second experiment (Section 6) were essentially the same as the present ones, except for the use of capitals.

issues, causing a bias towards the shortest alternative. Reference *d* and referent *r* were always in document parts whose layout properties differed from each other (e.g., *not* both in subsections labeled as "C" in different sections of the document). Had *d* and *r* occurred in document parts with similar layout properties, there might have been a bias towards the most complete (i.e., the longest) description.

5.2 Results

Hypotheses 1.1 and 1.2 were confirmed; hypothesis 1.3 was not. In fact, *DE* was avoided in 100% of all subjects' decisions. In situations involving *LO*, the *FI* version was chosen on average in 93% of cases (stdev = 15%), which is highly significant (Wilcoxon signed ranks test, $Z = -3.56$, $p < .0001$). In the cases not involving *DE* or *LO*, there was no significant preference for or against logical redundancy (Wilcoxon signed ranks test $Z = -0.51$, $p = .61$). The trend is in the predicted direction (mean of 57% for *MD* descriptions), but the variation between subjects was very large (stdev = 41%).

5.3 Discussion of First Experiment

This first experiment supported the hypothesis that subjects prefer references that include logically redundant information where there is a risk of *DE*/*LO*. Arguably, it is precisely this kind of information that is needed for the construction of NLG algorithms. Where logically redundant information does not make the referent easier to identify, the results of the experiment are less clear, with the subjects being divided between logically minimal and logically redundant descriptions. In other words, while supporting the informal observations reported in Sections 2 and 3, the experiment does not point to a generic preference of one of the two GRE algorithms presented in Section 4.

Evidently, there are many factors that this experiment did not address, such as the "distance" between objects. For example, if tables are enumerated throughout the document, is the brief, *SL*-type description *Table 5* easy enough to resolve? It depends: If there are tables on virtually every page then resolution is easy, because the table numbers support browsing not unlike page numbers; if tables are sparse, however, then searching through the entire document may take unacceptably long, and a more redundant, *FI*-type description such as *Table 5 in Section 4.3* is likely to be preferred. The nature of the domain is bound to matter as well. For example, in a large spatial domain in which navigation requires physical effort, short, *SL*-style descriptions are probably less acceptable than in a situation where the domain can be surveyed at a glance. To exemplify the first type of situation, let us return briefly to Examples (1a)–(1c), assuming that a city is divided into areas, and an area into streets:

(1a) 968 Lewes Road, Moulsecoomb area (FI-style)

(1b) 968 Lewes Road (SL-style)

(1c) number 968 (MD-style)

If these are uttered somewhere in Brighton but not on Lewes Road then *AS* predicts that utterance (1c) leads to *LO*, because the hearer will start looking for a number 968 in the street where the description is uttered. Consequently, utterance (1c) is infelicitous

anywhere except on Lewes Road. But how about Examples (1a) and (1b)? Both descriptions avoid *LO* and *DE*, because Brighton has only one Lewes Road. Yet if the hearer does not know that Lewes Road is in Moulsecoomb, then the resolution of Example (1b) may involve more work than Example (1a).

This experiment attempted to find out what types of references are favored by human judges when their opinion about these references is asked. Although this has the advantage that subjects were in a position to make trade-offs between the advantages and disadvantages of the different expressions (perhaps balancing ease of interpretation with ease of resolution), the method is limited in other respects. One limitation arises from the fact that meta-linguistic judgments are sometimes thought to be an unreliable predictor of people's linguistic behavior (e.g., van Deemter 2004). Perhaps more seriously, the experiment fails to tell us how difficult a given type of reference (for example, one of the *DE* type) would actually be for a reader, and whether the difficulty is a matter of interpretation or resolution. For these reasons, we decided to perform another experiment.

6. Second Experiment: Measuring Search Effort

In the previous experiment, we found that human authors often prefer logically redundant references, particularly when *DE* and *LO* can arise. In a follow-up experiment, we investigate the effect of logical redundancy on the performance of readers. We are primarily interested in understanding the search process, so resolution rather than interpretation. It will become clear that the new experiment necessitates a more careful design and a more complex analysis than the previous one.

6.1 Experiment Design

Subjects. Forty-two students on a first-year Computing Science course participated in the experiment as part of a scheduled practical.

Procedure. A within-subjects design was used. All subjects were shown 20 on-line documents. The order of the documents was randomized per subject, to control for order effects. The document structure was always visible, and so was the content of the current document part. A screenshot of an example document providing this level of information is shown in Figure 4. Each document was initially opened in Part B of either Section 2 or 3, where a task was given of the form "Let's talk about [topic]. Please click on [referring expression];" for instance: *Let's talk about elephants. Please click on picture 5 in part A.* Subjects could navigate through the document by clicking on the names of the parts (e.g. Part A as visible under Section 3). As soon as the subject had correctly clicked on the picture indicated, the next document was presented. Subjects were reminded throughout the document about the task to be accomplished, and the location at which the task was given. All navigation actions were recorded. At the start of the experiment, subjects were instructed to try to accomplish the task with a minimal number of navigation actions.

Reader's Knowledge. We assume that readers do not have *complete* knowledge of the domain. So, they do not know which pictures are present in each part of each section. If readers had complete knowledge, then a minimal description would suffice: For example, if readers knew that there is only one picture 5 in the document, located in

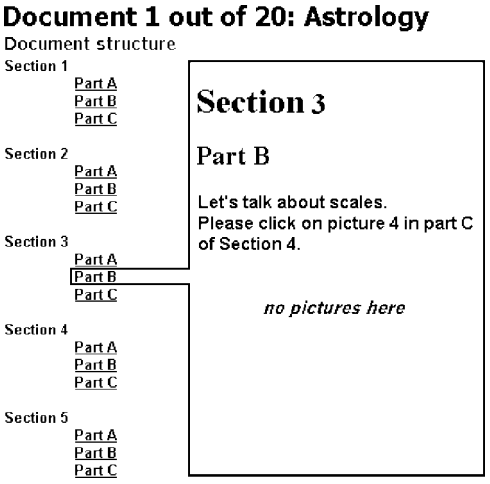

Figure 4
Fragment of the experiment interface.

part B of section 3, then the description *picture 5* would probably be completely clear. We do not, however, assume readers to be completely ignorant, either. We assume that they have *some* knowledge of the domain, particularly of its hierarchical structure. This brings us to the question of how much knowledge we should assume our readers to have. In practice (unlike Section 3.1, where the hearer was pictured as blindfolded until the description is uttered) readers will always have some knowledge: If in Part B of Section 2, then they would know (by convention) that there will also be a Section 1, and a Part A in Section 2, and so on. It is also likely that being in Part B of Section 2 and seeing pictures 1, 2, 3, readers will infer that sections can have parts, that parts can contain pictures, and that pictures are numbered (though not necessarily per part). Because of these kinds of consideration, it seems appropriate to give our readers knowledge about the entire document structure (the 5 sections and their parts) and the content (i.e., the existing pictures) in the current document part (but crucially, no knowledge about pictures elsewhere in the document, which require navigation to be discovered). A navigation structure like the one in Figure 4 provides this knowledge to the readers.

Research Questions. We want to test whether longer descriptions indeed help resolution, particularly in so-called problematic situations. Table 2 shows the types of situation (potential *DE*, *LO*, and non-problematic),[12] reader and referent location, and descriptions used.

> **Hypothesis 2.1:** In a problematic (*DE*/*LO*) situation, the number of navigation actions required for a long (*FI*/*SL*) description is smaller than that required for a short (*MD*) description.

12 In *DE* situations, there is another picture with the same number as the referent, but not in a part with the same name as the part in which the referent is. In *LO* situations, there is no other picture with the same number as the referent, and the reader location contains pictures. In non-problematic situations, there is another picture with the same number as the referent, but not in a part with the same name as the part in which the referent is.

Table 2
Situations of reference for Experiment 2.

Sit.	Type	Reader Loc.	Referent Loc.	Short (MD)	Long (FI/SL)	Long (other)
1	DE	Part B Sec 3	Part A Sec 2	*Pic 3 in Part A*	*Pic 3 in Part A Sec 2*	
2	DE	Part B Sec 2	Part C Sec 3	*Pic 4 in Part C*	*Pic 4 in Part C Sec 3*	
3	LO	Part B Sec 3	Part A Sec 3	*Pic 5*	*Pic 5 in Part A*	*Pic 5 in Part A Sec 3*
4	LO	Part B Sec 2	Part C Sec 2	*Pic 4*	*Pic 4 in Part C*	*Pic 4 in Part C Sec 2*
5	LO	Part B Sec 3	Part A Sec 4	*Pic 5*	*Pic 5 in Part A Sec 4*	*Pic 5 in Part A*
6	LO	Part B Sec 2	Part C Sec 1	*Pic 4*	*Pic 4 in Part C Sec 1*	*Pic 4 in Part C*
7	NONE	Part B Sec 2	Part A Sec 2	*Pic 3 in Part A*		*Pic 3 in Part A Sec 2*
8	NONE	Part B Sec 3	Part C Sec 3	*Pic 4 in Part C*		*Pic 4 in Part C Sec 3*

This hypothesis is similar to hypotheses 1.1 and 1.2 of the previous experiment. We will use the *DE* and *LO* situations in Table 2 to test this hypothesis, comparing for each situation the number of navigation actions of the short, that is, minimally distinguishing (*MD*) and long (*FI/SL*) expressions.

In the previous experiment, we had an additional hypothesis about non-problematic situations, stating that *MD* descriptions would be preferred to long descriptions in non-problematic situations. This is not a natural hypothesis in the new experiment, because it might not happen very often that a shorter description will lead to *fewer* navigation actions (**pace** Cremers 1996). (Note that in the previous experiment we looked at the combination of interpretation and resolution, whereas we are now focusing on resolution only). Instead, we will look at **gain**: the number of navigation actions required for a short description minus the number of navigation actions required for a long description.

> For situation s, short description sd of s, and long description ld of s, $Gain(s, sd, ld) =$ the number of navigation actions required in s for description sd minus the number of navigation actions required in s for description ld.

> **Hypothesis 2.2:** The gain achieved by a long description over an *MD* description will be larger in a problematic situation than in a non-problematic situation, that is, for problematic situation ps, non-problematic situation nps, *MD* description md of both ps and nps, and long description ld of ps and nps: $Gain(ps, md, ld) > Gain(nps, md, ld)$.

We will use the *DE* and non-problematic situations in Table 2 to test this hypothesis, comparing the gain of situation 1 with that of situation 7, and the gain of situation 2 with that of situation 8.

Longer descriptions may *always* lead to fewer navigation actions, and it can be expected that complete descriptions of the form *picture x in part y of section z* will outperform shorter descriptions in any situation. So, from a resolution point of view, an algorithm that would always give a complete description may produce better results

than the algorithms we proposed (e.g., situations 3 and 4 in Table 2). The aim of our algorithms is to make the descriptions complete enough to prevent *DE* and *LO* in *resolution*, but not overly redundant as this may affect *interpretation*. We would like to show that the decisions taken by *FI* and *SL* are sensible, that is, that they produce descriptions that are neither too short nor too long. Therefore:

> **S1:** We want to consider situations in which *FI* and *SL* have produced an incomplete description, and investigate how much gain could have been made by using a complete description in those cases. We would like this gain to be negligible. We will use situations 3 and 4 for this, calculating the gain of the long, complete descriptions (namely, **long (other)** in Table 2) over the shorter, incomplete descriptions generated by our algorithms (**long (*FI*/*SL*)** in Table 2).

> **S2:** We want to consider situations in which *FI* and *SL* have produced a complete description, and investigate how much gain has been made by using this compared to a less complete description that is still more complete than *MD*. We would like this gain to be large. We will use situations 5 and 6 for this, calculating the gain of the long complete descriptions generated by our algorithms (**long (*FI*/*SL*)** in Table 2) over the less complete descriptions (**long (other)**).

Introducing separate hypotheses for cases $S1$ and $S2$ poses the problem of defining when a gain is "negligible" and when a gain is "large." Instead, we will compare the gain achieved in $S1$ with the gain achieved in $S2$, expecting that the gain in $S2$ (which we believe to be large) will be larger than the gain in $S1$ (which we believe to be negligible).

> **Hypothesis 2.3:** The gain of a complete description over a less complete one will be larger for situations in which *FI* and *SL* generated the complete one, than for situations in which they generated the less complete one. More formally, for situations $S1$ and $S2$, descriptions *cd* and *ld*, with *cd* a complete description of $S1$ and $S2$ that has been generated by *FI* and *SL* for $S2$, and with *ld* an incomplete but longer-than-*MD* description of $S1$ and $S2$ that has been generated by *FI* and *SL* for $S1$: $Gain(S1, ld, cd) < Gain(S2, ld, cd)$.

Materials. Twenty on-line documents were produced,[13] with the same document structure (sections 1 to 5 with parts A to C) and containing 10 pictures. Documents had a unique background color, title, and pictures appropriate for the title. The number of pictures in a section or part varied per document. All of this was done to prevent subjects relying on memory. For instance, if we had used the same document for all tasks, subjects might have remembered where a particular picture was located. If we had used documents that looked similar, subjects might have assumed that they were the same. If we had kept the distribution of images the same, subjects might have learned that a particular part always contained many pictures.

Controlled experiments have advantages and disadvantages. Instead of using artificial, hand-crafted materials, we could have used real-world documents, like patient information leaflets, in order to make the tasks as realistic as possible. However, it would have been extremely difficult to find real-world documents that contain the right phenomena in a well-balanced way. Firstly, real documents might not have the right descriptions in them, so we would probably have needed to change sentences in the documents by hand. Secondly, we need a set of documents that are sufficiently

13 `http://www.csd.abdn.ac.uk/~jmasthof/RefStudy/Intro.php`.

Table 3
Number of clicks used to complete the tasks.

		Short		Long (FI/SL)		Long (Other)	
Sit.	Type	Mean	STDEV	Mean	STDEV	Mean	STDEV
1	DE	3.58	2.14	1.10	0.50		
2	DE	3.85	3.28	1.30	1.31		
3	LO	5.60	4.84	1.93	1.29	1.23	1.27
4	LO	2.50	1.97	1.60	1.28	1.38	2.07
5	LO	8.53	4.15	1.15	0.53	5.65	6.74
6	LO	7.38	5.49	1.25	1.03	4.08	2.35
7	NONE	1.58	0.98			1.63	2.61
8	NONE	1.48	0.96			1.05	0.32

similar in structure that one can make a fair comparison between longer and shorter descriptions. Moreover, the structure should not allow subjects to learn where in the document pictures are most likely to be located. Thirdly, semantic information or their background knowledge of the domain should be irrelevant. (For example, if we were using a real document on animals, and subjects read a section on lions, then they might expect a picture of a tiger to be in a nearby section, and a picture of an elephant to be closer than a picture of a pigeon.)

6.2 Results

Forty subjects completed the experiment. Table 3 shows descriptive statistics for the number of clicks subjects made to complete each task. To analyze the results with respect to Hypothesis 2.1, we used a General Linear Model (*GLM*) with repeated measures. We used two repeated factors: Situation (situations 1 to 6) and Description Length (short and long(*FI/SL*)). We found a highly significant effect of Description Length on the number of clicks used to complete the task ($F_{1,39} = 262.46$, $p < .001$, $\eta_p^2 = .87$). In all potentially problematic situations, the number of clicks is smaller for the long than for the short description. This confirms Hypothesis 2.1. We also found significant effects of Situation ($F_{5,35} = 13.11$, $p < .001$, $\eta_p^2 = .65$), and of the interaction between Situation and Description Length ($F_{5,35} = 18.02$, $p < .001$, $\eta_p^2 = .72$).

Table 4 shows descriptive statistics for the gain as used for Hypothesis 2.2. We again used a *GLM* with repeated measures, using two repeated factors: Description Content (that of situations 1 and 7, and that of situations 2 and 8) and Situation Type (potential *DE* and non-problematic).[14] We found a highly significant effect of Situation Type on the gain ($F_{1,39} = 26.62$, $p < .001$, $\eta_p^2 = .41$). In the non-problematic situations the gain is smaller than in the potential *DE* situations. This confirms Hypothesis 2.2.

Table 5 shows descriptive statistics for the gain as used for Hypothesis 2.3. We again used a *GLM* with repeated measures, using two repeated factors: Description Content (that of situations 3 and 5, and that of situations 4 and 6) and *FI* Decision (with 2 levels:

14 There were no significant effects of Description Content and of the interaction between Description Content and Situation Type. From here on, we will focus on effects that were significant.

Table 4
Gain as used for Hypothesis 2.2.

Sit.	Type	Mean	STDEV
1	DE	2.48	2.24
7	NONE	−0.05	2.77
2	DE	2.55	3.62
8	NONE	0.43	1.04

complete and not complete). We found a highly significant effect of *FI* Decision on the gain ($F_{1,39} = 24.10$, $p < .001$, $\eta_p^2 = .38$). The gain is smaller for situations where our algorithm decided to use an incomplete description than in situations where it chose a complete one. This confirms Hypothesis 2.3.

6.3 Discussion of Second Experiment

What does the second experiment teach us, over and above what we learned from the first one? First of all, the experiment suggests an explanation of why it was that, in problematic situations, subjects (in the first experiment) preferred redundant descriptions: The new experiment suggests that the reason may lie in the fact that, in the potentially problematic situations, the addition of structural information reduces the effort involved in resolution. This is, of course, exactly in line with the way in which *DE* and *LO* were introduced in Section 3, and with the assumptions about ease of resolution that were formulated in Paraboni and Van Deemter (2002a) and in the present Section 2.

Do our experiments, taken together, tell us *how much* redundancy is optimal in any given situation? In answering this question, let us first realize that pragmatic factors relating to the utterance situation are likely to affect how much redundancy is needed. At one end of the spectrum, there may be highly fault-critical settings, where flawless understanding is essential; at the other end, there may be discourse settings where accurate understanding is not important, and where the speaker/writer is under time pressure. Surely, redundant information must be more common in the former than in the latter. No one algorithm can cater to all types of settings.

On the other hand, our data do suggest quite strongly that, at least in the situation in which our subjects found themselves, a law of *diminishing returns* is in operation. To see this, let us first focus on the two *non-problematic* situations (Table 2): Averaging numbers

Table 5
Gain as used for Hypothesis 2.3.

Sit.	FI Decision	Mean	STDEV
3	NOT COMPLETE	0.70	1.40
5	COMPLETE	4.50	6.67
4	NOT COMPLETE	0.23	2.51
6	COMPLETE	2.83	2.16

of clicks of all subjects over all relevant situations, short descriptions required a mere 1.53 clicks; by adding redundant information (unlike *SL/FI*), this number gets reduced to an average of 1.34 clicks (long(other), in situations 7 and 8). This very slight gain (0.19 clicks) is not statistically significant ($F_{1,39} = .60$, $p = .44$, $\eta_p^2 = .02$) and is bought at the price of a description that is one and a half times longer, which makes it likely to take more time during *interpretation*. As for the more interesting *problematic* situations, perhaps the best comparison is between situations 3 and 4 (where long(other) exists and is longer than long(*FI/SL*)). Here, short descriptions lead to a pretty dismal average of 4.05 clicks. If we lengthen the descriptions as prescribed by *FI/SL* (long(*FI/SL*)) then this figure is lowered drastically to what looks like a pretty acceptable 1.77 clicks, which constitutes a gain of 2.28. By adding even more information (as in long(other)), the figure is lowered further, to 1.31 clicks. Although this does represent a gain, it is not statistically significant ($F_{1,39} = 2.94$, $p = .095$, $\eta_p^2 = .07$), and besides it is so small (at 0.46 clicks) that it seems likely to be more than offset by the disadvantages for *interpretation* that are implied by the increased length of the description. Needless to say, these effects can only become stronger if more complex documents are considered, and with descriptions that are even longer. Really excessive redundancy might have detrimental effects on resolution as well as interpretation, because it confuses hearers. (A hearer might wonder, along Gricean lines, "Why are they saying 'Picture 5 in Part A of Section 3, printed in black and white'. Surely if they have to give so much information, they cannot simply mean Picture 5?".)

Finally, we also explored the searching behavior of our subjects, focusing on the 12 documents in which incomplete descriptions were given. Ancestral Search predicts that subjects will search the current section (where the question is asked) exhaustively, before moving on to another section. Figure 5 shows subjects' compliance with Ancestral Search in their first navigation action. (Eight of the 12 documents contained a description of the form *Picture 5 in Part A*, so for these it suffices to look at the first navigation action.) Four subjects complied perfectly. Half the subjects complied almost perfectly, deviating in at most 2 of the 12 cases. However, five subjects deviated almost completely (10 or more times). Closer inspection showed that these latter subjects seemed to navigate randomly, not following any obvious pattern (e.g., top to bottom). It may well be that these subjects did not take the experiment seriously. Nevertheless, we still have more deviation from Ancestral Search than expected.

There are two possible explanations. First, some subjects may have started using Ancestral Search, and then found that it was not effective when they encountered some documents in which the referent turned out to be in some far-away section, after which they changed to a more random strategy. (Recall that our experiment deliberately

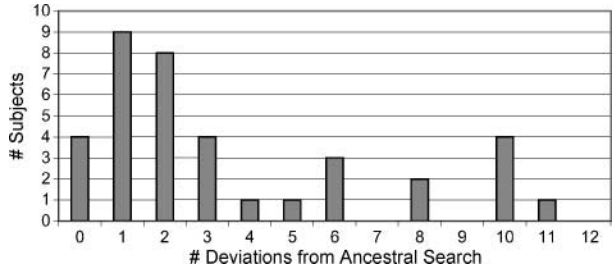

Figure 5
Compliance with Ancestral Search during first navigation action.

included some unreasonably short descriptions.) Our data seem to confirm this. For instance, subject *S*11 started in compliance with Ancestral Search until encountering a document asking, in Section 2, to find a picture in Part C. The subject clicked as many as 6 times on Part C of Section 2, before finally finding the referent in Section 3. He went on to deviate four times from Ancestral Search.

A second explanation for deviating from Ancestral Search is the kind of navigation that we allowed. Subjects could go directly from, say, Part C in Section 2, to part A in Section 3, without an extra navigation step to go into Section 3. In fact, it may even be *faster* to navigate to another section than within the current one, depending on the position of the mouse pointer. (This contrasts with the university domain, where one could not go directly from room 120 in Watts building to room 140 in Cockcroft building without first having to walk between the buildings.) It should be noted that this problem may be more pronounced after the first navigation action has been made. For instance, if one clicks on Part A in Section 2, then the mouse pointer is about as close to Part C in Section 1 as to Part C in Section 2. To explore this idea, we looked at the four documents in which a description of the form *picture 5* was given. In 83 cases, subjects who complied with Ancestral Search for the first navigation action needed to perform a second action; in 77% of these cases, they also complied with Ancestral Search in the second action. Now in as many as 68% of the cases in which they did *not* comply, they clicked on the closest link in an adjacent section (e.g., Part A of the next section after having first clicked on Part C). This confirms our suspicion that the lack of effort required to deviate may have been a reason for deviation. With hindsight, we should probably have made the distance between the relevant sections larger.

7. Conclusion

This article has discussed generation strategies that facilitate resolution of a referring expression by adding logically redundant properties. We have shown that this can be of crucial importance, especially in large domains, where minimally distinguishing descriptions can sometimes be completely useless (witness, e.g., Example [1c]). Two algorithms for generating logically redundant references along the lines described in this article have been implemented. The experiments reported in the previous sections indicate that these algorithms are fundamentally on the right track.

We recently learned of an interesting series of experiments that investigate the role of logically redundant properties in referring expressions (Arts 2004). One of the outcomes of these experiments was that certain types of logically redundant information almost consistently led to accelerated resolution. This was particularly true for information concerning the location of an object. For example, a logically minimal description like *the white button on the left* took readers longer to resolve than a redundant one like *the white button at the top left* (our emphasis). It is interesting to note that these results were obtained in situations where neither *LO* nor *DE* could occur.

This article has described an alternative to classical algorithms for GRE. Suppose you are designing an NLG system and want to give it a GRE component; how do you know whether to use the new algorithm, instead of one of its predecessors? Redundancy has a role to play in different kinds of situations (see the Introduction of this article), but our algorithms focus on a class of cases that we believe to be particularly widespread, namely where the domain is hierarchical in the sense of Section 3. Because hierarchies involve relations, let us once again compare the predictions made by our algorithms with those made by Dale and Haddock (1991). Suppose their description *the bowl on the*

table was said when there are two tables and two bowls, while (only) the table furthest away from the hearer has a bowl on it. *FI* and *SL*, by contrast, would generate something redundant like *the bowl on the far-away table.* Which of the two descriptions is best?

The answer is that it depends on the situation: When all the relevant facts are available to the reader without effort (e.g., all the domain objects are visible at a glance) then Dale and Haddock's minimal descriptions are fine, but when search is required, the kind of "studied" redundancy embodied in *FI* and *SL* becomes necessary. Consider the example again. If the tables and bowls are visible at a glance, then resolving the *DE*-inducing description *the bowl on the table* is unproblematic, because there is nothing here to discover: The crucial part of the domain is directly available, and no search is needed. Consequently, it is superfluous to say anything about the location of the table. But suppose we are in a huge room, where it is not obvious for the hearer what is on each table. In this situation, *the bowl on the table* would be a rather unhelpful description, compared to *the bowl on the far-away table* (or *the bowl on the table in the corner*), as would be consistent with our algorithms. (The example can be made more dramatic by hiding the table with the bowl on it in another room.) What this example highlights is the distinction between the things that speaker and hearer know when a referring expression is uttered, and the things they can *discover*. It is in the latter case that search becomes an issue. We have shown how this idea can be made precise and incorporated into a GRE algorithm, and we have demonstrated that this can improve the generated descriptions from the perspective of the hearer.

Recent work in psycholinguistics, focusing on spontaneous speech in dialogue, has shown that speakers and hearers often act as if they are completely oblivious of the epistemic limitations of their interlocutors, even when these limitations have been made perfectly obvious to them (e.g., Keysar, Lin, and Barr 2003). These widely known results have caused some researchers to expect language users to behave with unbridled descriptive "egocentricity" in all situations. The first of our two experiments suggests that human *writers* (as opposed, perhaps, to speakers) can be highly altruistic in their descriptions of objects. The second experiment demonstrates how descriptive altruism can benefit readers.

By exploring the benefits for the hearer (in terms of the effort required for identifying the referent), we have not only shown that it can be good to add logically redundant information to a referring expression; we have arguably also shed some light on the *reason* why redundant descriptions are sometimes preferred. By counting the number of clicks that subjects need in order to find the referent, and relating these to predictions stemming from our Ancestral Search model, we believe that we have achieved a degree of insight into the "resolution" processes in the head of the reader, not unlike the way in which insights in human language processing can be produced by *eye-tracking* experiments. It would be interesting to see whether the ideas discussed here can be confirmed using such a more entrenched psycholinguistic paradigm.

Acknowledgments

The authors are grateful for insightful comments from Emiel Krahmer, Richard Power, Sebastian Varges, the Aberdeen NLG group, and the anonymous reviewers. The second author acknowledges support from the UK's EPSRC TUNA project, grant GR/S13330/01.

References

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

Arts, Anja. 2004. *Overspecification in Instructive Texts.* Ph.D. thesis, Tilburg University, The Netherlands. Wolf Publishers, Nijmegen.

Association of the British Pharmaceutical Industry (ABPI). 1997. *1996–1997 ABPI Compendium of Patient Information Leaflets*. ABPI, London.
Clark, Herbert. 1992. *Arenas of Language Use*. CSLI Publications, Stanford, CA.
Cremers, Anita. 1996. *Reference to Objects; an Empirically Based Study of Task-oriented Dialogues*. PhD. thesis, University of Eindhoven.
Dale, Robert. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL-1989)*, pages 68–75, Vancouver, Canada.
Dale, Robert and Nicholas Haddock. 1991. Generating referring expressions involving relations. In *Proceedings of EACL-1991*, 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.
Deutsch, W. 1976. Sprachliche Redundanz und Objectidentifikation. Ph.D. dissertation, University of Marburg.
Edmonds, Philip G. 1994. Collaboration on reference to objects that are not mutually known. In *Proceedings of COLING-1994*, pages 1118–1122, Kyoto.
Grice, Herbert P. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics: Vol 3, Speech Acts*. Academic Press, New York, pages 43–58.
Horacek, Helmut. 2005. Generating referential descriptions under conditions of uncertainty. In *10th European Workshop on Natural Language Generation (ENLG-2005)*, pages 58–67, Aberdeen, Scotland.
Jordan, Pamela W. 2000. Can nominal expressions achieve multiple goals? An empirical study. In *ACL-2000*, pages 142–149, Hong Kong.
Jordan, Pamela W. 2002. Contextual influences on attribute selection for repeated descriptions. In K. van Deemter and R. Kibble, editors, *Information Sharing*. CSLI Publications, Stanford, CA, pages 295–328.
Keysar, Boaz, Shuhong Lin, and Dale J. Barr. 2003. Limits on theory of mind use in adults. *Cognition* 89:25–41.
Krahmer, Emiel and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing*. CSLI Publications, Stanford, CA, pages 223–264.
Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.
Mangold, Roland. 1986. *Sensorische Faktoren beim Verstehen ueberspezifizierter Objektbenennungen*. Peter Lang Verlag, Frankfurt.
Norman, Donald. 1988. *The Design of Everyday Things*. Doubleday, London.
Paraboni, Ivandré. 2000. An algorithm for generating document-deictic references. In *Proceedings of INLG-2000*, "Coherence in Generated Multimedia," pages 27–31, Mitzpe Ramon, Israel.
Paraboni, Ivandré. 2003. *Generating References in Hierarchical Domains: The Case of Document Deixis*. Ph.D thesis, University of Brighton, U.K.
Paraboni, Ivandré, Judith Masthoff, and Kees van Deemter. 2006. Overspecified reference in hierarchical domains: Measuring the benefits for readers. In *Proceedings of INLG-2006*, pages 55–62, Sydney.
Paraboni, Ivandré and Kees van Deemter. 2002a. Generating easy references: the case of document deixis. In *Proceedings of INLG-2002*, pages 113–119, New York.
Paraboni, Ivandré and Kees van Deemter. 2002b. Towards the generation of document-deictic references. In K. van Deemter and R. Kibble, editors, *Information Sharing*. CSLI Publications, Stanford, pages 329–354.
Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:98–110.
Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge.
Siddharthan, Advaith and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of 42nd ACL*, pages 408–415, Barcelona, Spain.
Sonnenschein, Susan. 1982. The effects of redundant communications on listeners: When more is less. *Child Development*, 53:717–729.
Sonnenschein, Susan. 1984. The effect of redundant communication on listeners: Why different types may have different effects. *Psycholinguistic Research*, 13:147–166.
van Deemter, Kees. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.
van Deemter, Kees. 2004. Finetuning an NLG system through experiments with human subjects: The case of vague descriptions. In *Proceedings of INLG-2004*, pages 31–40, Brockenhurst, UK.