Overspecified reference in hierarchical domains: measuring the benefits for readers

Ivandré Paraboni
University of Sao Paulo
EACH - Av.Arlindo Bettio, 1000
03828-000 Sao Paulo, Brazil
ivandre@usp.br

Judith Masthoff
University of Aberdeen
Dep.of Computing Science
Aberdeen AB24 3UE, Scotland, UK
jmasthoff@csd.abdn.ac.uk

Kees van Deemter
University of Aberdeen
Dep.of Computing Science
Aberdeen AB24 3UE, Scotland, UK
kvdeemte@csd.abdn.ac.uk

Abstract

It is often desirable that referring expressions be chosen in such a way that their referents are easy to identify. In this paper, we investigate to what extent identification becomes easier by the addition of logically redundant properties.We focus on hierarchically structured domains, whose content is not fully known to the reader when the referring expression is uttered.

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 1984, Clark 1992, Cremers 1996, Arts 2004, Paraboni and van Deemter 2002, van der Sluis, 2005). The present paper follows in the footsteps of Paraboni and van Deemter (2002) by focussing on hierarchically structured domains and asking whether any benefits are obtained when an algorithm for the generation of referring expressions (GRE) builds logical redundancy into the descriptions that it generates. Where Paraboni and van Deemter (2002) reported on the results of a simple experiment in which subjects were asked to say which description they *preferred* in a given context, the present paper describes a much more elaborate experiment, measuring how difficult it is for subjects to find the referent of a description.

1 Background

Let us 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 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 centerstage in our investigation.

Difficulty of resolution and interpretation do not always go hand in hand. Consider sentences (1a) and (1b), uttered somewhere in Brighton but not on Lewes Road.

(1a) *968 Lewes Road*
(1b) *number 968*

Assume that (1a) refers uniquely. If other streets in Brighton do not have numbers above 900, then even (1b) is a unique description – but a pretty useless one, since it does not help you to find the house unless your knowledge of Brighton is exceptional. The description in (1a) is longer (and might therefore take more time to read and interpret) than (1b), but the additional material in (1a) makes *resolution* easier once interpretation is successfully completed. We explore how an GRE program should make use of logically redundant properties so as to simplify resolution (i.e., the identification of the referent).

In corpus-based studies, it has been shown that logically redundant properties tend to be included when they fulfill one of a number of pragmatic functions, such as to indicate that a property is of particular importance to the speaker, or to highlight the speaker's awareness that the referent has the property in question (Jordan 2000). However, redundancy has been built into GRE algorithms

Proceedings of the Fourth International Natural Language Generation Conference, pages 55–62,
Sydney, July 2006.

only to a very limited extent. Perhaps the most interesting account of overspecification so far is the one proposed by Horacek (2005), where logically redundant properties enter the descriptions generated when the combined *certainty* of other properties falls short of what is contextually required. Uncertainty can arise, for example, if the hearer does not know about a property, or if she does not know whether it applies to the target referent.

Our own work explores the need for overspecification in situations where each of the properties in question is unproblematic (i.e., certain) in principle, but where the reader has to make an effort to discover their extension (i.e., what objects are truthfully described by the property). We ask how a generator can use logically redundant information to reduce the search space within which a reader has to 'find' a referent. (Cf., Edmonds 1994 for a related set of problems.)

2 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. While such simplifications permit the study of many aspects of reference, other aspects come to the fore when larger domains are considered.

Interesting questions arise, for example, when a large domain is hierarchically ordered. 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 belong to at most one of n's children, while 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, etc.; 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 favourite ways of structuring the world.

A crucial question, in all such cases, is what knowledge is shared between speaker and hearer at utterance time. It will be convenient to start by focussing on the extreme case where, before the start of resolution, 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

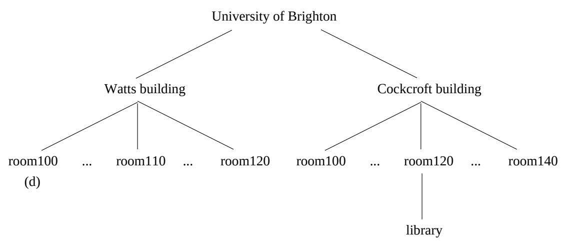

Figure 1: A hierarchically structured domain.

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, 'R.NUMBER=102' identifies a room uniquely within a given building) [1].

3 Obstacles for resolution

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 when the locations of speaker and hearer are taken into account. (For simplicity, we assume that the locations coincide.)

Suppose a hierarchically-ordered domain D contains *only one* entity whose TYPE is LIBRARY. Consider the following noun phrases, uttered in the position marked by d in Figure 1. (The first three have the same intended referent.)

[1]This is a useful assumption, since the existence of a distinguishing description cannot be otherwise guaranteed.

(2a) *the library, in room 120 in the Cockcroft bld.*
(2b) *the library, in room 120*
(2c) *the library*
(2d) *room 120*

Utterances like (2a) and (2b) make use of the hierarchical structure of the domain. Their content can be modelled as a list

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

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 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. For example, (2a) is modelled as

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

We focus on the search for x_n because, under the assumptions that were just made this is the only place where problems can occur (since no parent node is available).

Even though each of (2a)-(2d) succeeds in characterising their intended referent uniquely, some of these descriptions can be problematic for the hearer. One such problem occurs in (2d). The expression is logically sufficient. 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 this case we will speak of *Lack of Orientation* (LO).

Even more confusion might occur if another library was added to our example, e.g., in Watts 110, while the intended referent was kept constant. In this case, (2c) would fail to identify the referent, 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, while a is located in b. Such cases of mutual identification are unproblematic in small, transparent, domains where search is not an issue, but in large hierarchical domains, they are not. For, like (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, while not being the intended referent of the relevant part of the description (i.e., room 120 Cockcroft). Confusion could easily result. In cases like this, we speak of a *Dead End* (DE).

In section 5 we will present evidence suggesting that instances of Dead End and Lack of Orientation may disrupt search in a sufficiently large or complex domain. For a theoretical discussion we refer to Paraboni and van Deemter (2002).

4 Generation algorithms

What kinds of expression would existing GRE algorithms produce in the situations of interest? Since 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, while 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, incidentally, also y) uniquely, since only one value of x can be used to verify the three predicates; this justifies the description *'the bowl on the table'*. This situation can be 'translated' directly into our university domain. Consider Figure 2, with one additional library in room 110 of the Watts building. In this situation, the com-

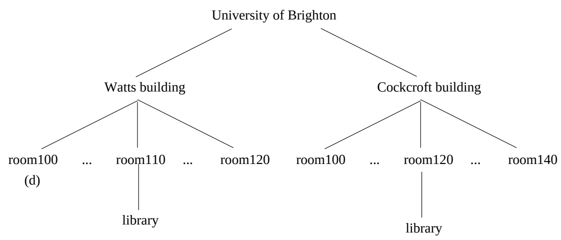

Figure 2: A university campus with two libraries.

bination $\{library(x), in(x, y), room(y), room-number(y) = 2\}$ identifies x (and, incidentally, also y) uniquely, because no other library is located in a room with 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.

Paraboni and van Deemter (2002) have sketched two GRE algorithms, both of which are guaranteed to prevent DE and LO by including logically redundant information into the generated descriptions so as to reduce the reader's search space. 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 materialise, and an intended referent r.

Briefly, the FI algorithm represents a straightforward way of reducing the length of search paths, without particular attention to DE or LO. 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 [2]. Applied to our earlier example of a reference to room 120, FI first builds up the list

$$L = \langle (r, \{type = room, r.number = 120\}) \rangle,$$

then expands it to

$$L = \langle (r, \{type = room, r.number = 120\}), \\ (x1, \{type = building, \\ buildingname = Cockcroft\}) \rangle.$$

Now that Parent(X) includes d , r has been identified uniquely within D and we reach STOP. L might be realised as e.g., 'room 120 in Cockcroft'.

FI gives maximal weight to ease of resolution. But something has to give, and that is brevity: By conveying logical redundancy, descriptions are lengthened, and this can have drawbacks. The second algorithm in Paraboni and van Deemter (2002), called SCOPE-LIMITED (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. This is done by making use of the notion of SCOPE, hence the name of the algorithm.

[2]The idea behind not moving up beyond this subtree is a natural extension of Krahmer and Theune's treatment of salience in GRE: see Paraboni and van Deemter (2002).

The *difference* between FI and SL becomes evident when we consider a case in which the minimally distinguishing description does not lead to DE or LO. For example, a reference to r = library would be realised by FI as 'the library in room 120 in Cockcroft'. By using SL, however, the same description would be realised by the SL algorithm simply as 'the library', since there is no risk of DE or LO. With the addition of a second library in the Watts building, the behaviour 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, SL would describe r as 'the library in room 120 of Cockcroft', just like FI. For details of both algorithms we refer to Paraboni and van Deemter (2002).

5 The new experiment

In Paraboni and van Deemter (2002) an experiment was described to find out what types of references are favoured by human judges when their opinion about these references is asked. As an example of a hierarchically ordered domain, the experiment made use of a document structured in sections and subsections. This allowed Paraboni and van Deemter (2002) to show their subjects the domain itself, rather than, for example, a pictorial representation (as it would be necessary in most other cases such as that of a University campus, which motivated many of our examples so far).

The experiment investigated the choice of so-called *document-deictic* references, such as 'the picture in part x of section y' made by authors of documents to check whether they choose to avoid potential DE and LO situations by adding redundant properties (favouring ease of resolution) and, conversely, whether they choose shorter descriptions when there is no such risk (favouring ease of interpretation). The results suggested that human authors often prefer logically redundant references, particularly when DE and LO can arise.

While this approach had the advantage that subjects could compare different expressions (perhaps balancing ease of interpretation with ease of resolution), the method is limited in other respects. For example, meta-linguistic judgements are sometimes thought to be an unreliable predictor of people's linguistic behaviour (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. Therefore, in this paper we report on a second experiment investigating the effect of the presence or absence of logical redundancy on the performance of readers. We are primarily interested in understanding the search process, so resolution rather than interpretation.

5.1 Experiment design

Subjects: Forty-two computing science students participated in the experiment, as part of a scheduled practical.

Procedure: A within-subjects design was used. Each subject was shown twenty on-line documents, in a random order. The entire 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 3. Each document was

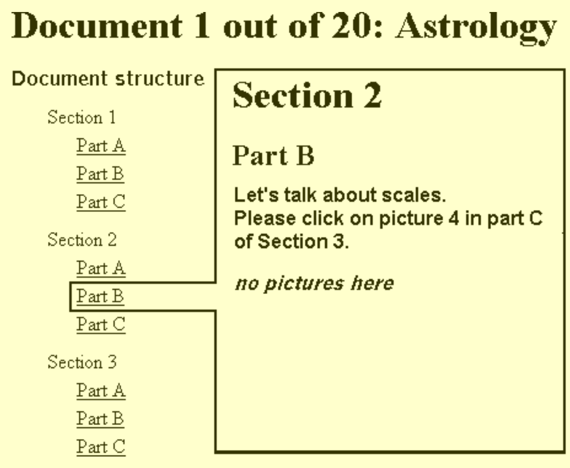

Figure 3: Fragment of the experiment interface.

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.

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. We do, however, not assume readers to be completely ignorant either[3]: we allowed them to see the current document part (where the question is asked) and its content, as well as the hierarchical structure (sections and parts) of the remainder of the document as in Figure 3 above.

Research Questions: We want to test whether longer descriptions indeed help resolution, particularly in so-called problematic situations. Table 1 shows the types of situation (potential DE, LO, and non-problematic)[4], reader and referent location, and descriptions used.

> **Hypothesis 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.

We will use the DE and LO situations in Table 1 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 Paraboni and van Deemter (2002) there was an additional hypothesis about non-problematic situations, stating that MD descriptions would be preferred to long descriptions in non-problematic situations. We cannot use this hypothesis in this experiment, as it is highly unlikely that a shorter description will lead to *fewer* navigation actions. (Note that the experiment in Paraboni and van Deemter (2002) looked at the combination of interpretation and resolution, while we are now focussing on resolution only). Instead, we will look at *gain*: the number of navigation actions required for a short description minus the number required for a long description.

[3]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 etc.

[4]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.

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

Table 1: Situations of reference

Hypothesis 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.

We will use the DE and non-problematic situations in Table 1 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, which do not always give complete descriptions (e.g. situation 3 in Table 1). 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, i.e. 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 1) over the short, incomplete descriptions generated by our algorithms (long (FI/SL) in Table 1).

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 1) over the less complete descriptions (long (other) in Table 1).

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 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.

Materials: Twenty on-line documents were produced, with the same document structure (sections 1 to 5 with parts A to C) and containing 10 pictures. Documents had a unique background colour, 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.

Documents were constructed specifically for the experiment. Using real-world documents might have made the tasks more realistic, but would have posed a number of problems. Firstly, documents needed to be similar enough in structure to allow a fair comparison between longer and shorter descriptions. However, the structure should not allow subjects to learn where pictures are likely to be (for instance, in patient information leaflets most pictures tend to be at the beginning). Secondly, the content of documents should not help subjects find a picture: e.g., if we were using a real document on animals, subjects might expect a picture of a tiger to be near to a picture of a lion. So,

		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

Table 2: Number of clicks used to complete the tasks.

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

Table 3: Gain as used for Hypothesis 2.

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

Table 4: Gain as used for Hypothesis 3.

we do not want subjects to use semantic information or their background knowledge of the domain. Thirdly, real documents might not have the right descriptions in them, so we would need to change their sentences by hand.

5.2 Results and discussion

Forty subjects completed the experiment. Table 2 shows descriptive statistics for the number of clicks subjects made to complete each task. To analyse the results with respect to Hypothesis 1, we used a General Linear Model (GLM) with repeated measures. We used two repeated factors: Situation (sit. 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 (p<.001). In all potential problematic situations the number of clicks is smaller for the long than for the short description. This confirms Hypothesis 1.

Table 3 shows descriptive statistics for the gain as used for Hypothesis 2. We again used a GLM with repeated measures, using two repeated factors: Descriptions Content (that of situations 1 and 7, and that of situations 2 and 8) and Situation Type (potential DE and non-problematic). We found a highly significant effect of Situation Type on the gain (p<.001). In the non-problematic situations the gain is smaller than in the potential DE situations. This confirms Hypothesis 2.

Table 4 shows descriptive statistics for the gain as used for Hypothesis 3. We again used a GLM with repeated measures, using two repeated factors: Descriptions Content (that of situations 3 and 5, and that of 4 and 6) and FI Decision (with 2 levels: complete and not complete). We found a highly significant effect of FI Decision on the gain (p<.001). The gain is smaller for situations were our algorithm decided to use an incomplete description than in situations were it chose a complete description. This confirms Hypothesis 3.

6 Conclusion

We have discussed generation strategies that facilitate resolution of referring expressions by adding logically redundant information to the descriptions generated. Redundancy has a role to play in different kinds of situation (see Introduction for references), but we have focussed on a class of cases that we believe to be widespread, namely where the domain is hierarchical. We have argued that, in such situations, minimally distinguishing descriptions can sometimes be useless. Various algorithms for generating logically redundant references have been implemented. The extensive experiment of section 5 indicates that these algorithms are fundamentally on the right track.

The new algorithms discussed in this paper are an alternative to classical GRE algorithms. This raises the question how one knows whether to use the new FI or SL instead of one of its competitors? Let us 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. In this situation, FI and SL would generate something redundant like *the bowl on the far-away table*. Which of the two descriptions is best? We submit that it depends on the situation: when all the relevant facts are available to the hearer without effort (e.g., all the domain objects are visible at a glance) then minimal descriptions are fine. But in a huge room, where it is not obvious to the hearer what is on each table, search is required. It is this type of situation that there is a need for the kind of 'studied' redundancy embodied in FI and SL, because the minimally *'the bowl on the table'* would not be very helpful. The new algorithms are designed for situations where the hearer may have to make an effort to uncover the relevant facts.

By focussing on the benefits for the reader (in terms of the effort required for identifying the referent), we have not only substantiated the claims in Paraboni and van Deemter (2002), to the effect that it can be good to add logically redundant information to a referring expression; we have also been able to shed light on the *reason* why redundant descriptions are sometimes preferred (compared with the experiment in Paraboni and van Deemter (2002), which did not shed light on the reason for this preference): we can now say with some confidence that, in the circumstances specified, the generated redundant descriptions are resolved with particular ease. By counting the number of clicks that subjects need to find the referent, we believe that we may have achieved a degree of insight into the 'resolution' processes in the head of the reader, not unlike the insights coming out of the kind of eye-tracking experiments that have been popular in psycholinguistics for a number of years now. It would be interesting to see whether our ideas can be confirmed using such a more entrenched experimental paradigm.

7 References

Arts, Anja. 2004. *Overspecification in instructive texts.* PhD. Tilburg University, The Netherlands. Wolf Publishers, Nijmegen.

Cremers, Anita. 1996. *Reference to Objects; an empirically based study of task-oriented dialogues.* Ph.D. thesis, University of Eindhoven.

Dale, Robert and Nicholas Haddock. 1991. Generating Referring Expressions involving Relations. EACL, Berlin, pp.161-166.

Dale, Robert and Ehud Reiter. 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science* 18:pp.233-263.

Deutsch, W. 1976. "Sprachliche Redundanz und Objectidentifikation." Unpublished PhD dissertation, University of Marburg.

Edmonds, Philip G. 1994. Collaboration on reference to objects that are not mutually known. COLING-1994, Kyoto, pp.1118-1122.

Krahmer, E. and Theune, M. 2002. Efficient Context-Sensitive Generation of Referring Expressions. In K. van Deemter and R. Kibble (eds.) *Information Sharing.* CSLI Publ., Stanford.

Horacek, Helmut. 2005. Generating referential descriptions under conditions of uncertainty. 10th European workshop on Natural Language Generation (ENLG-2005). Aberdeen, pp.58-67.

Jordan, Pamela W. 2000. Can Nominal Expressions Achieve Multiple Goals?: An Empirical Study. ACL-2000, Hong Kong.

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

Mangold, Roland. 1986. *Sensorische Faktoren beim Verstehen ueberspezifizierter Objektbenennungen.* Frankfurt: Peter Lang Verlag.

Paraboni, Ivandre. 2000. An algorithm for generating document-deictic references. INLG-2000 Workshop Coherence in Generated Multimedia, Mitzpe Ramon, pp.27-31.

Paraboni, Ivandre and van Deemter, K. (2002). Generating Easy References: the Case of Document Deixis. INLG-2002, New York, pp.113-119.

Sonnenschein, Susan. 1984. The effect of redundant communication on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research* 13, pp.147-166.

van Deemter, Kees. 2004. Finetuning an NLG system through experiments with human subjects: the case of vague descriptions. INLG-04, Brockenhurst, UK, pp.31-40.

van der Sluis, I. 2005. Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions. Ph.D. thesis, Tilburg University, the Netherlands.