THE GENERATION OF NATURAL DESCRIPTIONS

CORPUS-BASED INVESTIGATIONS OF REFERRING EXPRESSIONS IN VISUAL DOMAINS

Henriette Anna Elisabeth Viethen

This dissertation is presented for the degree of

Doctor of Philosophy

at

March 2011

Declaration

The research presented in this thesis is the original work of the author except where otherwise indicated. Some parts of the thesis include revised versions of published papers. This work has not been submitted for a degree or any other qualification to any other university or institution. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged. The research presented in this thesis was approved by the Macquarie University Ethics Review Committee, reference number: `HE27FEB2009-D06283` on 16 January 2009.

Signed: Henriette Anna Elisabeth Viethen

Date:

Abstract

Referring expression generation (REG) has been studied by computational linguists for nearly three decades. Although other aspects of the task have been examined, most investigations into REG are focussed on the selection of those attributes of an object that best distinguish it from all others in its environment. Historically, much of this work has suffered from two problems: firstly, it does not take account of empirical evidence for how people refer; and secondly, it has not been evaluated against human-produced corpora.

This thesis is based on two related premises which I take to be self-evident if our ultimate goal is to explain how humans refer: first, that naturalness should be the primary goal of computational models of referring expression generation, and second, that the task therefore needs to be approached by using human-produced corpora for the development and testing of algorithms.

Based on these premises, this thesis presents an extensive exploration into how corpora can be used in REG. It makes three main contributions in this area: (1) it presents a study that explores how corpora can be used to evaluate algorithms for the generation of referring expressions, and shows that existing algorithms cannot fully account for the way humans generate referring expressions; (2) it provides a detailed analysis of the different aspects of the human use of referring expressions in two large corpora in order to inform the development of REG algorithms; and (3) it presents experiments in using these corpora to train decision trees for attribute selection for referring expressions. The main conclusion of the analyses and experiments in this thesis is that speaker-specific variation plays a much larger role in the generation of referring expressions than existing algorithms acknowledge.

Chapter 2 begins by surveying existing research in the field of REG. Chapter 3 then provides an in-depth discussion of the methodological choices that have to be made when employing corpora to inform and evaluate REG algorithms. Chapter 4 presents an evaluation of three popular existing REG algorithms using a small corpus of human-produced data. It shows that, while one of the algorithms is capable of generating a large proportion of the referring expressions in the corpus, none of them are even in principle able to generate all of them. The experiment gives rise to a dissection of the issues involved in the evaluation of REG algorithms. Based on the analyses of the previous three chapters, Chapter 5 describes the design, collection and annotation of two large corpora of referring expressions, and analyses how speakers make use of different object properties. These corpora are novel in that they contain spatial relations between objects, allowing a systematic analysis of the circumstances under which people use relations as well as other properties. The second corpus constitutes the largest systematically-designed single-domain collection of referring expressions to date. Finally, Chapter 6 explores the use of the corpora described in Chapter 5 to train algorithms which model the content selection behaviour of the human participants who contributed the data. Modelling this data using decision trees is a natural way to gain insights into the factors that influence a person's decision to include a particular property in a referring expression and how these factors interact.

Acknowledgements

First of all, I thank my supervisor Robert Dale. I can truly say that without him this thesis would never have happened. I thank him for believing in me and in my project, even when I didn't, for encouraging me to try things out and for forcing me to finish things off, for sending me to every conference and every summer school, for teaching me how to write research papers, for showing me what it means to treat students with respect, for never putting his own interests before mine, and for forgiving me my inability to let go of German punctuation rules.

I thank Albert Gatt, Alexander Koller, Anja Belz, Ehud Reiter, Emiel Krahmer, Ielka van der Sluis, Imtiaz Khan, Kees van Deemter, Mariët Theune, and Ross Turner for interesting discussions and for reminding me that there are other people interested in the same things as me, albeit mostly on the other side of the world. I thank my three thesis reviewers, John Kelleher, Kathy McCoy, and Kees van Deemter for their constructive comments and suggestions. Special thanks go to Meg Mitchell who probably has no idea that in a short online chat she explained to me what my work was all about. Without her I might never have figured out how to pull it all together into one rounded piece of work.

I thank the Markists and the CLT lunch gang for their support and friendship and for making lunch breaks worth the daily trek to uni: Andrew Lampert, Ben Hachey, Ben Phelan, Diego Molla, Elena Akhmatova, François Lareau, Jean-Philippe Prost, Jojo Wong, Luiz Pizzato, Marc Tilbrook, Mark Dras, Mark Johnson, Mary Gardiner, Matt Honnibal, Menno van Zanen, Pawel Mazur, Rolf Schwitter, Stephen Wan, Suzy Howlett, Teresa Lynn, Yasaman Motazedi, various visitors over the years, and, in particular, Simon Zwarts, who patiently helped me with countless trivial technical problems and programming questions without ever being condescending and who just generally is a great friend.

I thank Ana Castro, Angélica Tomaz, Bonne Eggleston, Duncan Macinnis, Emily Mitchell, Geoff Thilo, Henry 'Woody' Woodruff, Jonathan Berant, Matt Roberts, Reut Tsarfaty, Rui Costa, Sascha Jockel, Shelley Bambrook, Susana Correia, Tanja Döring, Telia Curtis and Yael Augarten, as well as Jorge Cham and his phdcomics, for constantly reminding me that I was not the only one fighting a thesis monster and that it is possible to finally slay it. I am especially grateful to Yael Augarten, Telia Curtis and the rest of the crew from G12B at the School for Photovoltaics and Renewable Energy at UNSW, who provided me refuge in their office, where I hid from the world (and the internet) for four months to write up the first draft of this thesis.

I thank my parents, Maria Viethen, Richard Eßer and Volkhart Schönberg, my brother Lasse, and the rest of my family back in Germany and Austria, who despite having lost me to Australia always welcome me back with open arms and have been supporting my PhD adventure from afar.

Most importantly, I thank Smithy for keeping me alive with his love and his amazing cooking skills. Without him, it would all amount to nothing.

Contents

Notational Conventions

Example	Description
the blue ball	Linguistic examples in the body of the text are in italics.
$E = \sum_{i=0}^{n} x$	Mathematical symbols and variable names are in italics.
cluster	Boldface is used for technical terms when they are first introduced.
Question 1	Capitalised boldface is used to highlight hypotheses and research questions in the running text.
ACL	Small capitals are used for acronyms and names, such as abbreviations for systems and algorithms.
tg_size	Lowercase sans serif terms are used in knowledge representation contexts.
TG_Size	Capitalised sans serif terms are used for machine learning features.

A note on the 'academic plural'

This thesis is written in the first person singular. Although many doctoral and masters theses are written in the 'academic plural', I feel at odds with this tradition of writing a monograph in the first person plural. This does not mean that I never use the first person plural. I do so when I refer to work that I have published in a co-authored paper, and in cases in which I invite the reader to join me in considering a particular idea or a certain section, table or figure of the thesis.

Chapter 1

Introduction

1.1 Problem Statement

When we talk or write, we constantly have to build referring expressions that describe the things we want to talk or write about. In order to ensure a successful discourse, these referring expressions have to be chosen in a way that allows the listener or reader to identify the correct thing, the intended referent. As humans, we do this effortlessly and most of the time without even thinking about it at all. It does not feel like a difficult task; it simply happens as one of the many background processes that support our language production facility.

The research field of Natural Language Generation (NLG) is concerned with the development of computational systems that generate natural language text. Of course, in order to generate texts that are useful to humans, such systems need to refer to things in a natural, human-understandable way, so building referring expressions is a crucial subtask of NLG. As with so many cognitive tasks that humans perform every day without much effort, getting a computer program to generate adequate referring expressions is not as straightforward. Over the last three decades the problem of referring expression generation (REG) has attracted more attention from computational linguists than most other aspects of NLG and has developed into a small research field of its own.

This level of attention is due in large part to the consensus view that has arisen as to what is involved in computational referring expression generation: the task is widely accepted to involve a process of selecting those attributes of an intended referent that distinguish it from other potential distractors in a given context. The bulk of the work in the field, and especially more recently developed algorithms, therefore primarily concentrate on the generation of context-free descriptions of objects with the single goal of identifying the target referent (cf., Dale and Reiter,

1995; van Deemter, 2000; Krahmer et al., 2003; Gardent, 2002; Horacek, 2003; Kelleher and Kruijff, 2006; Gatt, 2007). Referring expressions of this kind are often referred to as **distinguishing descriptions**. A few exceptions exist in approaches which are capable of taking into account at least some of the discourse context around the referring expression under construction (Dale, 1990; Heeman, 1991; Edmonds, 1994; Gardent and Striegnitz, 2007; Jordan and Walker, 2005).

Many authors have pointed out the importance of generating referring expressions that sound as much like human-produced descriptions as possible (cf. Gardent et al., 2004; Horacek, 2004; van der Sluis and Krahmer, 2004a; Gatt, 2007; Gatt et al., 2007), and three recent shared-task evaluation competitions (STECs) in this field have put much emphasis on evaluating the output of candidate systems for human-likeness. Striving for human-likeness in the generation of referring expressions serves two purposes. First, models capable of mimicking human behaviour can claim at least some level of cognitive plausibility and bring us closer to being able to understand and explain human behaviour, thereby contributing to the aims of cognitive science. Second, by building REG models on the basis of such an understanding of human behaviour we can ensure that their output sounds natural to humans and fulfils the task it is aimed at without being confusing or carrying false implicatures: a model that can emulate speakers' reference behaviour also emulates speakers' ability to suit the needs of listeners. Human–human communication is for the large part extremely successful, which means that systems that do what people do have a good chance of also satisfying the needs of their users.

However, much existing work has focussed primarily on other concerns, by taking account of issues regarding computational complexity and by concentrating on the production of descriptions which are in some sense minimal, in that they do not contain unnecessary information (Dale, 1989; Dale and Haddock, 1991a; Dale and Reiter, 1995; Gardent, 2002; Horacek, 2003; Areces et al., 2008). At the time that they were conceived, the traditional algorithms for referring expression generation found in the literature were not based on empirical evidence for how people actually refer; nor was their behaviour evaluated against human-produced referring expressions by the original authors. A small number of recent evaluation exercises as well as a short series of STECs on the task of content determination for REG constitute evidence that this trend is about to turn (Gupta and Stent, 2005; Gatt et al., 2007; Viethen et al., 2008; Belz and Gatt, 2007; Gatt et al., 2008, 2009a).

The key premise of this thesis is that human-likeness has not occupied the central position in referring expression generation that it should, and that the other criteria that have been considered in order to determine what counts as a

good algorithm have been given undue weight. This premise is based on three observations.

First, it is clear (and this observation is not new) that humans do not always produce what are referred to as minimal distinguishing descriptions (i.e. referring expressions whose content walks the line between being both necessary and sufficient), despite this having served as a concern for much algorithmic development in the past. As has long been recognised, human-produced referring expressions are in many cases informationally redundant. The Incremental Algorithm (Dale and Reiter, 1995), which serves as the basis for many algorithmic developments in the literature, is occasionally given credit because it can lead to referring expressions that contain redundancy; but even its authors were careful not to claim that the redundancy it produces is the same as that produced by humans. The kinds of redundancy generated by algorithms have never been compared to those evident in human-produced referring expressions, and this has led to systems which at best pay lip-service to the need to account for redundancy.

Second, it can be expected that the first practical NLG applications of the algorithms developed in the area of REG will be those where the location of objects within physical scenes is a requirement: examples of such scenarios are the description of entities such as buildings and other landmarks in automatically generated route descriptions (see Dale et al., 2005) and the description of locations in 'omniscient room' scenarios, where an intelligent agent might try to tell you under which sofa cushion you left your RFID-tagged keys. In these scenarios, it is very likely that any referring expression generated will need to make use of the spatial relationships that hold between the intended referent and other entities in the domain to be of any use to a listener; but, surprisingly, the generation of relational references is a relatively unexplored task. The few algorithms that address this task (Dale and Haddock, 1991a; Gardent, 2002; Krahmer and Theune, 2002; van der Sluis and Krahmer, 2005; Kelleher and Kruijff, 2006) typically favour fairly simple approaches: they only consider spatial relations if it is not possible to fully distinguish the target referent from the surrounding objects in any other way, or they treat them in exactly the same way as non-relational properties, which leads to awkward sounding expressions that humans would neither produce nor easily be able to understand. None of these approaches would generate the relational referring expressions that I introduce in this thesis.

The third observation (also not particularly new, but surprisingly ignored by most REG algorithms) is that different people do different things when faced with the same reference task. This poses serious questions for both the development of algorithms and their evaluation: as has been noted for other tasks that involve

natural language output (such as document summarisation and machine translation), in such circumstances we clearly cannot evaluate an algorithm by comparing its results against a single gold-standard answer. Even with a range of possible candidate answers, it is still possible that an algorithm might produce a perfectly acceptable solution that is not present amongst this set. This forces us to consider more carefully what it is that we are doing when we develop algorithms for the generation of referring expressions (or, for that matter, for any generation task): are we trying to emulate or predict the behaviour of individual speakers in a given situation? Or are we trying to produce a solution which might somehow rate as optimal in a task-based evaluation scenario (such as might be measured by the time it takes a listener to locate a referred-to object), recognising that human-produced referring expressions are not necessarily optimal in this sense? In this thesis I take the first of these two perspectives, with the aim of shedding some light on the different strategies that people use when they produce referring expressions.

1.2 Contributions

This thesis takes the view that the best way to advance the field of REG at its current stage is by studying the way humans produce referring expressions. By gaining a more detailed understanding of the processes that humans apply to this task, we can increase our chances of success at building REG algorithms that produce truly natural-sounding and useful referring expressions. In order to study the way people build distinguishing descriptions, we need large corpora of such descriptions that can both inform algorithm development and be used for the evaluation of human-likeness.

Therefore, rather than developing a rule-based REG algorithm in the tradition of existing approaches, this thesis presents an extensive exploration of the ways in which corpora can be used in the field of referring expression generation. It makes three main contributions in this area:

- It presents a study that explores how corpora can be used to evaluate algorithms for the generation of referring expressions. In this study, three of the classical REG algorithms were re-implemented and their output compared to the referring expressions contained in a collection of human-produced descriptions in a simple domain of drawers in filing cabinets. The main result of this pilot evaluation demonstrates that the existing algorithms cannot fully account for the way humans generate referring expressions, in particular those containing spatial relations and redundant information. The study serves as a platform for the discussion of a number of problems inherent in

corpus-based evaluation of REG and possible solutions for these problems. Here, I also examine how these issues have impacted on the recent shared task evaluation challenges in REG and point to some problems with the way human-likeness of REG algorithms was evaluated in these challenges.

- I introduce two new corpora of referring expressions. I describe the details of two data collection experiments and the analysis of a number of different aspects of the human use of referring expressions in the resulting corpora. The main focus of the analysis is on the use of spatial relations between objects, with the aim of gaining a better understanding of human reference behaviour in this respect. This focus on spatial relations is a direct consequence of the results of the above-mentioned evaluation experiment, which found that existing algorithms are particularly bad at replicating the use of spatial relations by humans. Both corpora contain data from many different participants for each stimulus item, which allows the exploration of cross-participant variation.

- The thesis presents experiments in using the new corpora to train decision trees for the task of attribute selection for referring expression generation. Using machine learning techniques to automatically extract behavioural patterns from data is the most direct way to ensure that the resulting systems are based on human behaviour. Additionally, decision trees make it relatively easy to inspect and interpret the resulting models, thus allowing insights into the factors that are influential. The main conclusions of these machine learning experiments are (1) that speaker-specific variation plays a much larger role in the generation of referring expressions than existing algorithms acknowledge, and (2) that by focussing on the individual attributes, rather than complete referring expressions, more commonality between speakers can be found.

1.3 Overview

In Chapter 2, I first define the task of referring expression generation more formally than I have done above, and then discuss the literature that forms the background for the research described in this thesis. Particular attention is paid to the generation of referring expressions containing relations and to existing work using corpora.

Chapter 3 sets up some methodological machinery that is needed for corpus-based work in REG, including corpus collection and analysis and corpus-based evaluation. It weighs up the advantages and disadvantages of the different options

that are available for collecting corpora from the viewpoints of naturalness and experimental control. Following this, the chapter gives an overview of qualitative differences in visual attributes of objects and discusses what it means for a referring expression to be minimal or over-specified. As we will see, these concepts are important for the analysis of corpora of referring expressions. Finally, it discusses different metrics that have been used to compare system output to human-produced data in REG and related fields.

In Chapter 4, I consider in more detail how REG algorithms can be evaluated against collections of human-produced referring expressions. These considerations are based on an evaluation experiment involving three classic REG algorithms and a small set of referring expressions. The direct results of this experiment show that two of the main challenges still facing REG are the generation of human-like redundancy and the production of referring expressions that use spatial relations between objects to identify the target referent. Based on the experience gained in this experiment, the second half of this chapter comprises an in-depth discussion of the problems that arise in corpus-based evaluation. These problems include issues such as determining input representation, dealing with the non-deterministic way in which natural language is used by humans, and finding adequate performance measures. Here, I also consider the way in which the recent community-wide STECs have handled these problems.

Two new collections of referring expressions are introduced in Chapter 5. The design of these corpora is based on an appraisal of the requirements arising from the discussion and experimental results of the previous chapters, in particular, the need to account for cross-speaker variation and the need for an investigation into the use of spatial relations. I describe the collection experiments that resulted in the two corpora and the analysis of the semantic content of the referring expressions they contain, with a focus on the use of spatial relations between objects.

In Chapter 6, the two corpora from Chapter 5 are used in a series of machine learning experiments. I trained decision tree models to predict the use of the different object attributes that occur in the corpora. The performance of the decision trees is used to uncover the differences in predictability of the use of different attributes and to determine which features of the referential scenarios contribute most to the characterisation of the participants' use of the attributes. Based on the results of these experiments, I make a case for **speaker profiles**, participant-specific sets of attribute-based decision trees, which capture commonalities in cross-participant behaviour at the level of individual attributes while accounting for much of the variation at the level of complete referring expressions.

Finally in Chapter 7, I summarise the outcomes of the research presented in this thesis and discuss ideas for future work that arise from it.

Chapter 2

Content Selection for Distinguishing Descriptions

In this chapter I survey the literature on referring expression generation from a mostly computational perspective. I attempt to provide a complete picture of research on REG within the Natural Language Generation community over the past two decades, which includes some influential work that is not directly relevant to the work presented in the later chapters; however, only work that has a bearing on the remainder of the thesis is described in full detail. The principal purpose of this review is to give an overview over the state of the art in REG, paying attention, in particular, to work that deals with relations between objects, work that makes use of corpora to inform or evaluate REG algorithms, and the different perspectives that work in REG has taken on what it means for a referring expression to be optimal.

First, I formally define what I take to be the task of a REG algorithm for the purpose of this thesis. Following this, I give an overview of the different computational models that have been proposed for REG and extensions to these models that allow a wider variety of phenomena in referring expressions to be generated. A separate section is dedicated to approaches that are able to generate referring expressions describing the target referent in terms of a relation to another object. Following this, I turn to corpus-based work in REG including empirical approaches and evaluation studies; and finally, I discuss different views on the optimality of referring expression.

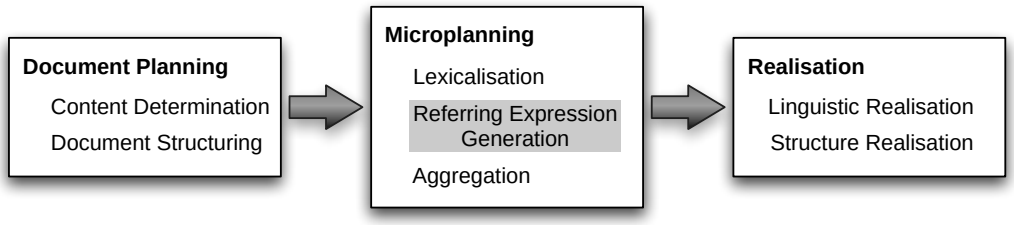

Figure 2.1: Reiter and Dale's (2000) pipeline architecture for NLG.

2.1 The Task of REG

The generation of referring expressions is one of the most studied problems within NLG. This is due both to the fact that it is perceived as one of the core tasks that every NLG system needs to tackle and to the fact that it is the one problem within NLG that is most clearly defined.

Reiter and Dale (2000) divide the task of NLG into a pipelined architecture of subcomponents. The pipeline is pictured in Figure 2.1. It consists of three stages, Document Planning, Microplanning and Realisation, each with its own subtasks. In this view, REG is one of the tasks of the pipeline taking place during the Microplanning stage of the NLG pipeline. However, when we consider REG in isolation, it quickly becomes clear that things are more complicated than this. It appears that some of the other subtasks within Reiter and Dale's pipeline are in actual fact tasks that have to be tackled within REG itself as well. A system charged with generating a fully-fledged natural language description will as a minimum have to perform the tasks of content determination (selecting the properties of the target referent to be mentioned), lexicalisation (choosing the words to represent the properties), and linguistic realisation (constructing a grammatically correct noun phrase). Even the task of aggregation can be argued to form part of the generation of a distinguishing description, as the semantic content might in some cases best be spread across several partially distinguishing noun phrases (Horacek, 2004). This suggests that rather than being treated as a component of a pipeline, the planning and realisation of referring expressions should be seen as interacting with the generation of the whole text and its parts at every level of processing.

As long as we concentrate on REG in isolation, this interaction with the generation of the context of a referring expression does not have to concern us. Instead we can concentrate on one or more of the subtasks of REG itself. By far the majority of work in REG concentrates on **content determination**, and this thesis is no

exception to this rule. Therefore, the definition of the task of referring expression generation that I adopt is solely concerned with the task of choosing the semantic content for a referring expression and assumes that lexical choice and surface realisation are tackled later on in the NLG architecture.[1]

Furthermore, I concentrate in this thesis on **distinguishing descriptions**, or **referential descriptions** as Donnellan (1966) calls them, that is on referring expressions whose main function it is to distinguish the target referent from the objects around it. Other functions that referring expressions can fulfil, such as informing, describing, convincing or directing, are beyond the scope of this work. Therefore, the term **referring expression generation** (or REG) will, with very few exceptions, be used in this thesis as a synonym for **content determination for identification**.

The above discussion puts me into a position to now define more explicitly the task that a referring expression generation algorithm is set to tackle and the terms relevant to it in the form that will be applicable to the discussions in the remainder of this thesis.

The **domain** defines the types of entities that are being referred to, in some cases even a particular set of entities with all their properties. For example, a domain might consist of a collection of real-world buildings, traffic lights, trees and other landmarks in their actual physical environment, if we are giving directions; or it might consist of a collection of photographic portraits which are displayed in 2D grids for experimental reasons; or it might consist of a collection of food ingredients that are not visually available. The term 'domain' is overloaded in much of the REG literature, being used both in the sense in which I have just defined it and in the sense of a context set, which I define as follows.

A **context set** contains a subset of the domain entities, for example, the landmarks visible at a certain point of the path for which we are giving directions, a subset of the photographs used in an experimental setup, or the cooking ingredients that have already been mentioned in a recipe. One of these entities is the **intended referent** or **target referent** r, the entity that is being referred to, and the remaining objects constitute the **distractor set** D, the objects from which r has to be distinguished. For each object x, the **knowledge base** KB holds a set of properties $\mathsf{P_x}$ that are true of this object. A **property** is an attribute–value pair of the form $\langle \mathsf{attribute{:}value} \rangle$ or a relation to another object: $\langle \mathsf{relation{:}object} \rangle$. For simplicity, I will sometimes refer to a property as a literal p only. In particular in

[1] This is not to say that I do not acknowledge that intricate interactions between surface realisation and content selection can exist and in some cases make it impossible to separate the two tasks.

the schematic displays of algorithms, I will take $\mathsf{p}(\mathsf{x}) \in \mathsf{KB}$ to mean that according to the knowledge base the object x has the property p.

In a given context set, each property of r has a certain **discriminatory power** which is defined as the proportion of the number of distractor objects that do not have this property (or the same value for the same attribute) to the total number of distractors (Dale, 1989). In a visual environment, the context set, including the properties of its member objects and their spatial configuration, is often called a **scene**.

Sometimes more information is available than what is contained in the context set or scene. For instance, we might have information about the speaker, if only in form of an ID from a data collection experiment; we might know how many referring expressions the speaker has already produced; or in a discourse context we might know whether the target referent has been referred to before and how. All of this information combined with the context set make up the **referential scenario**. In many cases in REG, the referential scenario is identical to the context set, because no additional information about the circumstances under which a reference occurred is known or deemed to be important.

In an experimental setting, the term **trial** denotes either a context set or a referential scenario, depending on the focus of the experiment. I will avoid the term in situations where this distinction is important.

Given an intended referent, a referential scenario and a knowledge base, the task of a referring expression generation algorithm is to find a set of properties of r that successfully distinguish it from all objects in D. It is such a set of properties that I call a **referring expression** or alternatively a **(distinguishing) description** L. More formally (following Reiter, 1990a), a referring expression is a distinguishing description if the following conditions hold:

- Each property in L is true of the target referent r.
- For each distractor d in D, there is at least one property p in L that does not apply to d.

In a corpus of referring expressions, the combination of a referential scenario with a referring expression produced in that scenario constitutes an **instance** of reference in this corpus. A corpus might contain several instances for the same context set and even for the same referential scenario. Some instances for a given referential scenario might result in the same referring expression, but this does not have to be the case.

2.2 REG Frameworks

Existing approaches to automatically determining the content for a referring expression can be classed roughly into two groups: those that propose new knowledge representation and reasoning mechanisms, and those that extend existing algorithms in order to broaden their coverage in terms of the phenomena found in human-produced referring expressions. In this section, I introduce the different reasoning frameworks that have been proposed for REG and a number of their extensions. Approaches to relational referring expressions are covered in Section 2.3.

2.2.1 Greedy Search

In his EPICURE system, Dale (1989, 1990, 1992) proposed the first fully formalised computational approach to content selection for referring expressions. Dale's algorithm performs a search over the set of properties of the target referent to choose the smallest possible subset that fully distinguishes the target from all distractor objects. Dale applied his algorithm in a domain of cooking recipes in which the context set was not visually available; rather all entities that had been introduced into the discourse and not yet destroyed (e.g. by chopping or mixing into a new entity such as a dough) were assumed to be part of the current context set. In this domain the algorithm was intended for subsequent reference; however, the same algorithm can be used without change for visual domains and for initial reference.

The algorithm attempts to balance three principles derived from Grice's (1975) conversational maxims:

1. **The Principle of Adequacy** states that a referring expression must contain enough information to distinguish the target referent unambiguously from all distractors and enable the listener to resolve the reference.

2. **The Principle of Efficiency** requires that a referring expression should not contain more information than necessary for identification, as that might result in unintended implicatures for the listener.

3. **The Principle of Sensitivity** specifies that a referring expression should be sensitive to the needs and abilities of the listener by only including information that he knows or can easily perceive.

As a result of the first two principles, the aim of this algorithm is **full brevity** (FB) (a term coined by Reiter, 1990a) which means finding the shortest possible distinguishing description for the target referent. Reiter (1990a) also pointed out that guaranteeing full brevity is a computationally intractable problem (it is

r — the intended referent
$\mathsf{D} = \{\mathsf{d} \mid \mathsf{d}$ *is a distractor of* $\mathsf{r}\}$
KB — the knowledge base containing the properties of all domain objects
$\mathsf{P_r} = \{\mathsf{p} \mid \mathsf{p}(\mathsf{r}) \in \mathsf{KB}\}$
$\mathsf{L} = \{\}$ — the empty description

1. Check Success:
   if $\mathsf{D} = \emptyset$ **then** return L as a distinguishing description
   elseif $\mathsf{P_r} = \emptyset$ **then** fail
   else goto Step 2
2. Choose Property:
   for each $\mathsf{p_i} \in \mathsf{P_r}$ **do**:
   $\mathsf{D_i} \leftarrow \mathsf{D} \cap \{x | x \in \mathsf{D}$ *and* $\mathsf{p_i}(x) \in \mathsf{KB}\}$
   Choose property $\mathsf{p_j}$ such that $\mathsf{D_j}$ is the smallest set
   goto Step 3
3. Extend Description (with the chosen $\mathsf{p_j}$):
   $\mathsf{L} \leftarrow \mathsf{L} \cup \{\mathsf{p_j}\}$
   $\mathsf{D} \leftarrow \mathsf{D_j}$
   $P \leftarrow P - \{\mathsf{p_j}\}$
   goto Step 1

Algorithm 2.1: **The** GREEDY **Algorithm (Dale, 1989)**

equivalent to the NP-hard set cover problem) and that the algorithm proposed by Dale (1989) in fact employs a greedy heuristic which can in some cases miss the shortest possible description. As a result, Dale's algorithm is usually referred to as the **Greedy Algorithm** (I will use the term GREEDY for short).

Pseudocode for GREEDY (adapted from Dale and Haddock, 1991b) is given in Algorithm 2.1. Given are a domain containing the intended referent r and a set of distractors D, a knowledge base KB containing the properties of all distractors, a separate set of properties $\mathsf{P_r}$ true of r, and the initially empty description L. In Step 1, the algorithm checks if it is done. This is the case if there are no more distractors left ($|\mathrm{D}| = \emptyset$) or if the algorithm runs out of properties for r ($\mathsf{P_r} = \emptyset$), in which case no distinguishing description can be found. In Step 2, the algorithm chooses that property of r that excludes most distractors from D and then in Step 3 adds this property to the description L and adjusts the distractor set D and the property set $\mathsf{P_r}$.

As mentioned above, the GREEDY algorithm in the EPICURE system was intended for subsequent reference. For this reason, the system was also capable of producing pronouns and one-anaphors, an aspect that has been much neglected in the later literature, which is more focussed on initial or one-off reference.

Although it has been superseded by the hugely popular Incremental Algorithm (see Section 2.2.2), many ideas and concepts from GREEDY permeate through much

of the subsequent work on the generation of distinguishing descriptions. Most notably:

- It is driven by Grice's (1975) conversational maxims and the avoidance of false implicatures.
- Properties are considered in **serial dependency**. In each iteration only one property is considered and its inclusion depends on the other properties that have already been chosen: a property's discriminatory power depends on the remaining distractor set which in turn is determined by the properties included so far.
- The underlying mechanism is a search over the properties of the intended referent.

The last of these points was further developed by Bohnet and Dale (2005) who couched a number of REG algorithms in a unified search framework based on the problem solving algorithm for AI proposed by (Russell and Norvig, 2003). They defined a REG problem as a state consisting of the description L constructed so far, the set of distractors D that L might refer to, and the set of properties $\mathsf{P_r}$ that have not been included in L yet. In the initial state, L is empty, D contains all other objects in the domain and $\mathsf{P_r}$ contains all properties known to be true of r. In the goal state, L contains some of the properties from $\mathsf{P_r}$ and D is empty, which means that L only refers to r. In Bohnet and Dale's model the search proceeds from state to state by moving properties between P and L and adjusting D accordingly. The search process and its outcome can be influenced by means of a cost function and different ways of enqueuing the properties in P and expanding the search graph.

2.2.2 Incremental Reference Generation

The Incremental Algorithm

As a response to the computational complexity issues involved in aspiring to produce the shortest possible distinguishing description, Dale and Reiter (Reiter and Dale, 1992; Dale and Reiter, 1995) proposed the Incremental Algorithm (IA). They sacrificed the goal of finding the shortest referring expression and thereby achieved polynomial time-complexity. They did this by simplifying the computationally expensive second step of GREEDY (see Algorithm 2.2). Instead of checking through the complete list of r's properties in each iteration and computing which one rules out most distractors, the IA simply chooses the first property it finds that rules out *any* distractors. It introduces the notion of a predefined **preference ordering** which determines in which order properties are considered for inclusion in the

r — the intended referent
$D = \{d \mid d \textit{ is a distractor of } r\}$
KB — the knowledge base
$P_r = \langle p \mid p(r) \in KB \rangle$ — ordered according to preference
$L = \{\}$ — the empty description

1. Check Success:
   if $D = \emptyset$ **then** return L as a distinguishing description
   elseif $P_r = \emptyset$ **then** fail
   else goto Step 2
2. Choose Property:
   for each $p_i \in P_r$ **do**:
   if $\{x | x \in D \textit{ and } p_i(x) \in KB\} \neq \emptyset$
   then choose p_j and **goto** Step 3
   else $P_r = P_r - p_j$
   goto Step 1
3. Extend Description (with the chosen p_j):
   $L \leftarrow L \cup \{p_j\}$
   $D \leftarrow D_j$
   $P \leftarrow P - p_j$
   goto Step 1

Algorithm 2.2: **The Incremental Algorithm (IA, Dale and Reiter, 1995)**

referring expression under construction. The authors intended this property ordering to allow a certain degree of context-sensitivity: for each new domain a new appropriate preference order could be defined. We will see in Chapter 4 that the preference order can have a very strong influence on the outcome, but that one preference order alone cannot account for all referring expressions. The preference order makes the IA much more adaptable than GREEDY but also introduces the non-trivial problem of having to find the best preference order for a given situation.

The incremental procedure results in an algorithm that may produce a description with properties which are strictly speaking informationally redundant. Although each property rules out at least one distractor at the time it is chosen, the lack of backtracking means that properties chosen later in the process can render already included ones redundant. The authors justified this move away from Grice's Maxim of Brevity and Dale's Principle of Efficiency with the observation that many human-produced descriptions are, in fact, over-specified and by basing their algorithm on the incremental model of language production (Schriefers and Pechmann, 1988; Levelt, 1989; Pechmann, 1989). This means that the IA might be able to produce more natural referring expressions than GREEDY. I will return to the topics of naturalness and over-specification in Sections 2.5.3 and 3.2.2, respectively, and both GREEDY's and the IA's ability to produce natural referring

expressions will be tested in Chapter 4.

In order to simplify comparisons to GREEDY and other algorithms I will review later in this chapter, the schema displayed in Algorithm 2.2 leaves out some of the detail of Dale and Reiter's original algorithm. None of the features left out here are important in the experiments presented in the following chapters. Dale and Reiter included a UserKnows function which prevents the algorithm from including any properties which the listener cannot be assumed to know about (e.g. because they have not been mentioned yet or are not visible). They also distinguished different levels of specificity for the values of each attribute and included a mechanism to choose the most appropriate level of specificity (findBestValue). For example, if the domain contains objects in several different shades of pink, it makes sense to use values such as hot pink or magenta to differentiate between them; however, if there is only one magenta object and a lot of green objects, calling the magenta one pink might be more appropriate. Finally, they treated the type property separately from the other properties, which allows them to ensure that (1) it is always included and (2) the most basic level class possible is used. The first of these requirements is based on the need for a property that can be realised as a head noun. I will discuss this requirement in Section 3.2.2. The second requirement follows psycholinguistic findings that for most entities a **basic-level class** exists and that using a more specific class label can confuse the listener in the same way as an overly redundant description might.

Extensions of the IA

Due to its computational efficiency and simplicity as well as its alleged psycholinguistic plausibility, the Incremental Algorithm has become the most implemented and built upon REG algorithm in the literature. In the following I describe some of these extensions of the IA.

Theune and Krahmer proposed an extension that allows the generation of subsequent reference with the IA taking into account the discourse salience of the target referent (Krahmer and Theune, 1998; Theune, 2000; Krahmer and Theune, 2002), and a second one which allows the IA to produce referring expressions that contain binary relations to other objects (Theune, 2000; Krahmer and Theune, 2002). I will return to their relational extension in Section 2.3. Theune and Krahmer's approach works by assigning a salience score to all objects according to the focus/topic distinction by Hajičová (1993) and Centering Theory (Grosz et al., 1995). They alter the success criterion of the algorithm and only let it stop when

there is no distractor left that is as or more salient than the target referent.

Not all properties are the same. The qualitative differences that exist between different properties were first discussed in the REG literature by van Deemter (2000, 2006). He pointed out that the appropriateness of **vague** or **gradable** properties such as small and large is dependent on the context in which they are used, while, for example, the colour of an object is absolute. Consider two descriptions in a domain of animals:

(2.1) the large animal

(2.2) the large mouse

In Description (2.1) the referent can be assumed to be large compared to all other animals. However, in Description (2.2) it is only likely to be large in comparison to other mice, not for example, to any elephants that might be present. This shows that the meaning of *large* in these examples is dependent on the type of the animal as well as the other animals around it. In fact a related observation was made in a psycholinguistic experiment by (Brown-Schmidt and Tanenhaus, 2006): size tends to be used only if there is another object in the domain which has the same type as the target referent, but is of different size. van Deemter proposed to deal with size properties in the IA by replacing absolute values in the knowledge base by a number of derived inequalities which compare the size of each object to the absolute sizes of all others. He then ensured that size always appears after all other properties in the preference order and that larger inequalities appear before smaller ones, based on the assumption that the meaning of vague terms is usually dependent on that of the other properties in the referring expression in which they occur. He then added a post-processing step that infers from the inequalities in a referring expression whether it is possible to use large or small or if the absolute value should be used.

In (van Deemter, 2002), van Deemter considered the IA's logical completeness in terms of the Boolean operators of negation and disjunction. He extended it to be able to generate referring expressions that contain negated properties, such as Example (2.3), and descriptions of sets of objects, such as Example (2.4), or even (2.5), which contains a logical disjunction of properties. His algorithm proceeds in stages, trying longer and longer disjunctions of properties, if atomic properties and shorter disjunctions did not suffice to distinguish the target set.

(2.3) the black dog that is *not* a poodle

(2.4) the black dogs

(2.5) the black dog *and* the poodle

The work on reference to sets was taken further by Gatt and van Deemter (2005, 2006), who have presented the most mature algorithms in this space to date. They used a similar procedure to the IA in that their algorithms are based on incremental processing of a preference order of properties. Their algorithms add a lot of complex machinery to the basic procedure to ensure that properties are chosen in a way that maximises coherence within the set of objects described by the referring expressions. For example, their approach will attempt to use properties of the same type for all the referents of a set. So, it would produce descriptions such as Examples (2.6) or (2.7) rather than Example (2.8) or (2.9)

(2.6) the musician and the professor

(2.7) the Italian and the Swede

(2.8) the Italian and the professor

(2.9) the musician and the Swede

Siddharthan and Copestake (2004, 2007) used the IA for the re-generation of referring expressions in newspaper text taken from the Penn Treebank. Their approach differs from other approaches to REG in that it is aimed at being useful for applications such as summarisation or question-answering and therefore takes text as input rather than a well-defined knowledge base. Working with words on a lexical level rather than with properties at a semantic level, they incorporated a refined version of discriminatory power into the IA which they call the **discriminating quotient**. In their approach, the preference order is sorted in such a way that properties that are less similar to those of the distractors according to WordNet (Miller et al., 1993) are considered before those that are similar to distractor properties. In (Siddharthan and Copestake, 2007) they tried a version of the IA in which the preference order is adjusted dynamically at runtime, listing the properties according to their discriminatory quotient at that point. This results in a similar selection behaviour to that of GREEDY. Siddharthan and Copestake (2004) also suggested an approach to discourse salience in the IA that is slightly different to that of Krahmer and Theune mentioned above. They proposed taking the salience of the distractor objects into account in the mechanism that determines the preference order: a property that distinguishes the referent from a salient distractor is worth more than one that distinguishes it from one with relatively low salience.

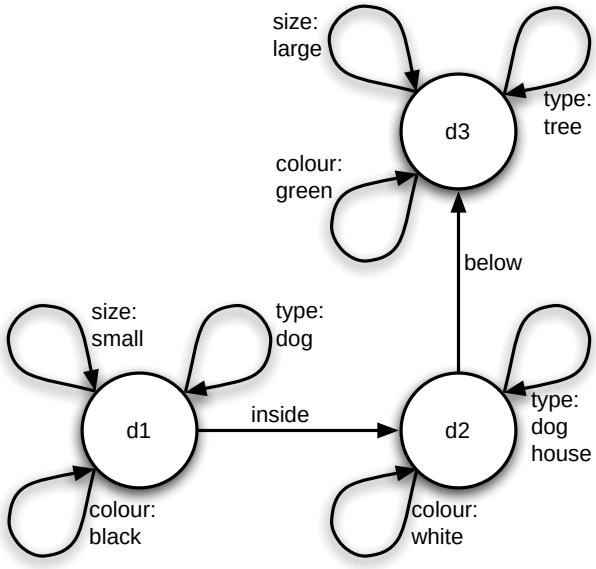

Figure 2.2: A sample domain represented as a labelled directed graph.

2.2.3 Graph-Based REG

The Graph-Based Framework

The approaches to REG discussed above all operate on essentially the same type of knowledge representation: a knowledge base containing the set of all entities in the domain and for each entity a set of properties that are true of it, and possibly a set of spatial relations in which it takes part. Sometimes this is varied by instead listing for each property which entities have this property. In either case this amounts to a propositional database containing a list of ⟨entity:attribute:value⟩ triples for atomic properties and ⟨relation:entity:entity⟩ triples for binary relations stored in one way or another.

The first approach to propose an improvement to the way the underlying knowledge base is represented, rather than the way search is performed on this knowledge base, was Krahmer et al.'s (2003) graph-based framework. They reformulated the task of selecting attributes for referring expressions as a graph-theoretical problem. To this end, the domain including the target referent and distractor objects is represented as a labelled directed graph. The graph representation of a visual scene models each object of the scene as a vertex in the **scene graph**. Atomic attributes such as colour, type or size are represented as looping edges on the corresponding node. They are labelled with the attribute names and the values the object in question has for these properties. Relations between objects, for example below or

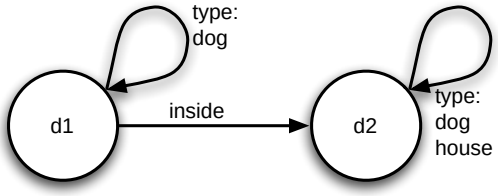

Figure 2.3: A sample description represented as a labelled directed graph.

inside, are modelled as edges between the corresponding vertices. Figure 2.2 shows a sample scene graph containing three objects: a dog, a dog house and a tree.

To generate a distinguishing description, the graph-based algorithm searches for a subgraph of the scene graph that uniquely identifies the target referent, called a **distinguishing graph**. Starting with the subgraph only containing the vertex which represents the target referent, it performs a breadth-first search over the edges connected to the subgraph found so far. It searches the space exhaustively, but uses a cost-based heuristic (described below) to effectively prune the search space.

Informally, a subgraph refers to the target referent if and only if it can be 'placed over' the domain graph in such a way that the subgraph vertex representing the target object can be 'placed over' the vertex of the target in the domain graph, and each of the labelled edges in the subgraph can be 'placed over' a corresponding edge in the domain graph with the same label and same direction. Furthermore, a subgraph is distinguishing if and only if it can be 'placed over' exactly one vertex in the domain graph. The informal notion of one graph being 'placed over' another corresponds to the mathematical graph-theoretic concept of **subgraph isomorphism**.

An example for a distinguishing subgraph describing d_1 in the sample domain graph in Figure 2.2 would be Figure 2.3, which could be realised as Description (2.10).

(2.10) the dog inside the doghouse.

Of course, *the dog* or *the small dog* would in this situation also suffice as distinguishing descriptions.

As mentioned above, the graph-based framework uses a cost function to guide the search through the space of possible subgraphs. This cost function is defined over all edge labels and vertices in the domain graph. The cost of a subgraph is then defined as the sum over all edges and vertices contained in it. The search algorithm is guaranteed to find the cheapest subgraph representing a distinguish-

ing description for the target referent. To avoid an exhaustive brute-force search through the entire space of subgraphs, the cost function is used to prune a search branch as soon as it becomes as or more expensive than the cheapest distinguishing subgraph found so far.

Using a cost function as a means to indicate a preference for certain properties over others makes it possible to specify the extent of the preference as well as equal preference for certain properties or even property values. Let us assume our target object is a friendly, small, white poodle and two possible distinguishing descriptions for it are:

(2.11) the friendly poodle [poodle, friendly]

(2.12) the small white one [white, small]

If the property costs are

c(poodle) = 1,
c(white) = c(small) = 11, and
c(friendly) = 12,

then poodle is very much preferred over the other properties, and white and small are equally preferred. For this cost function, the algorithm will choose Description (2.11) with cost 13 over Description (2.12) with cost 22, although both properties appearing in Description (2.12), white and small, are preferred over friendly which appears in Description (2.11). If the cost distribution is instead

c(poodle) = 1,
c(white) = c(small) = 3, and
c(friendly) = 12,

then Description (2.12) with cost 6 is chosen over Description (2.11), costing 13, although poodle in Description (2.11) is preferred over both white and small in Description (2.12).

This kind of choice is not possible in the IA which only uses a preference ordering over the properties. The IA's preference order is a simple ranking and does not indicate how much one property is preferred over others. It therefore does not make it possible to specify whether the preference for poodle over white and small outweighs the preference for white and small over friendly.

Of course, it is still possible that more than one distinguishing subgraph with the lowest cost exists. In this case the subgraph encountered first will be the one returned by the algorithm as the description for the target referent. The order in

which subgraphs (i.e. descriptions) are found is dependent on the order in which edges (i.e. properties) were considered during the search process. This means the earlier a property is considered, the more likely it is going to be part of the referring expression produced when there is more than one cheapest solution. In other words, in addition to the cost function, the graph-based framework is also controlled by a preference order.

Extensions of the Graph-Based Framework

Multimodality is an aspect of language use that has not been focussed on in the REG literature much. One notable exception is the work of van der Sluis and Krahmer (Krahmer and van der Sluis, 2003; van der Sluis, 2005), who proposed an extension of the graph-based framework that integrates pointing gestures into referring expressions. Their approach is based on psycholinguistic experiments investigating the way that people combine gestural and verbal information when referring. It represents pointing gestures of differing preciseness as loop edges on the target referent in the domain graph. The more imprecise the pointing gesture, the more of the target's closest neighbours are included in the gesture and therefore have the same pointing edge. The cost of pointing edges is determined by the size of the target referent and the distance that the pointing device has to travel to make the pointing gesture associated with it. The more imprecise the pointing gesture, the less effort is involved in bringing the pointing device into the correct position and therefore the cheaper the pointing gesture.

Based on the observation that the verbal information in such multimodal referring expressions is often redundant when the pointing gesture is taken into consideration, van der Sluis and Krahmer (2005) also proposed an adaptation of their algorithm that allows the generation of over-specified multimodal descriptions. They made use of a certainty score that represents the speaker's estimate of how likely the listener is to misinterpret the referring expression under construction. The certainty score of a property is determined by a hierarchy over the domain attributes, whereby absolute attributes, such as colour, have a higher certainty score than relative ones such as size. The certainty score of a pointing gesture is dependent on its preciseness and calculated in a similar way to the cost of pointing gestures. The certainty score of a referring expression is the sum of the certainty scores of the properties and pointing gestures contained in it. As long as the overall score is below a certain threshold, more properties have to be included, even if the description is already distinguishing from a logical point of view. This is reminiscent of Edmonds' (1994) **confidence threshold**, which determines

whether the speaker thinks that the referring expressions is salient enough for the listener to be able to resolve it successfully.

Many of the extensions that were proposed for the IA can be adapted to the graph-based framework as well, as was argued by van Deemter and Krahmer (2007). They showed that simple reference to sets and gradable properties can be treated in the graph-based framework in the same way as was proposed by van Deemter (2000) for the IA. They also demonstrated that Krahmer and Theune's (2002) approach to discourse salience in the IA can easily be implemented in the graph-based freamwork by restricting the distractor set to those entities that are at least as salient as the target referent. Additionally, van Deemter and Krahmer illustrated how negated properties could be integrated by making them explicit in the domain graph, under the closed world assumption that every property that is not true in the knowledge base is false. However, this approach might require some fine-tuning. In their example domain, descriptions such as Examples (2.13) and (2.14) could be produced, where Example (2.14) refers to a person.

(2.13) the trumpet not holding the small musician

(2.14) the small non-trumpet

These two examples demonstrate that a stricter control for what is physically possible, or likely, needs to be applied when the domain graph is extended. For example, the negation of a relation should only be applied to entities of the same type as those between which the positive relation holds. Negations of types might either best be avoided all together or only applied to entities with a type from the same type hierarchy. For example, it might make sense under some circumstances to describe an animal as *the only non-mammal* where animal and mammal are clearly from the same type hierarchy, while describing a person as *the thing that is not a trumpet* seems to be a much less acceptable description.

Conceptual Graphs

Conceptual Graphs (CGs) are a logic-based knowledge representation formalism proposed by Sowa (1984) which can be mapped to First-Order Logic formulas. A CG consists of a bipartite graph, which specifies the factual knowledge about a domain and is comparable to Krahmer et al.'s domain graph, and a number of support hierarchies which capture ontological knowledge about the types of entities and relations that can be involved in the domain graphs. The domain graphs contain two different kinds of nodes: entities are represented as rectangles

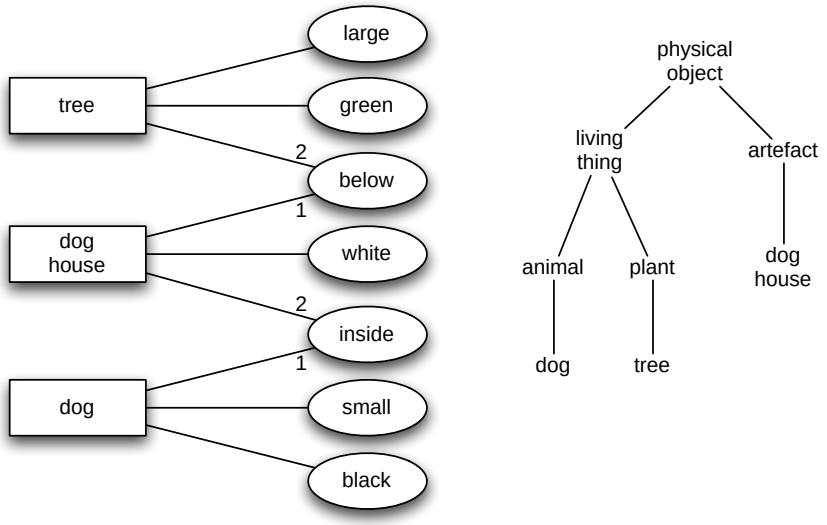

Figure 2.4: The domain from Fig. 2.2 as a Conceptual Graph.

and relations between entities are represented as circles. Edges only exist between entity and relation nodes, which makes the graphs bipartite. The edges are labelled with numbers to indicate the directionality of relations and the number of edges connected to a given relation node defines the relation's arity. Atomic properties are represented as unary relations connected to only one entity node. Figure 2.4 shows a CG representation of the example domain from Figure 2.2.

Croitoru and van Deemter (2007) proposed using CGs as the representational formalism for referring expression generation in combination with a simple breadth-first search algorithm. They show that, while existing approaches can be represented in the CG framework, it offers a number of advantages that go beyond the capabilities other frameworks. Firstly, CGs are firmly anchored in First-Order Logic, which allows tapping into established computational mechanisms and simplifies the consideration of complexity and expressivity of different algorithms. Secondly, binary relations are handled in the same way as atomic properties, and relations of any higher arity can be represented and treated in a natural way without complex additional machinery or peculiar adaptations of the knowledge representation. Thirdly, ontological background information is already part of the formalism in the form of the support hierarchies over concepts and relations. This makes the integration of functions such as `FindBestValue` and `BasicLevelValue` proposed by Dale and Reiter (1995) straightforward.

2.2.4 Other Frameworks

(Dale and Haddock, 1991b) introduced a constraint-based approach to REG for their extension of GREEDY to relational descriptions. The sets of properties already chosen for each of the objects that have been introduced into the description constitute the set of constraints. Every time a new property is included into this constraint set, the set of distractors for each of the objects in the description is adjusted accordingly. I will describe this algorithm in more detail in Section 2.3. Gardent and colleagues (Amoia et al., 2002; Gardent, 2002; Gardent et al., 2004) use a similar constraint-based approach, which can also be used for the generation of plural descriptions.

Varges (2004) introduced a chart-based overgenerate-and-rank approach to REG which separates the representational form of the knowledge base from that of the referring expression being built. This allows the algorithm to logically infer information that is not represented explicitly in the knowledge base, rather than adding such information to the knowledge representation, as was done in extensions to the IA and the graph-based framework described above. Referring expressions containing Boolean combinations of properties as well as relations between entities fall out naturally from Varges' approach. His algorithm first builds up all possible combinations of properties to describe all objects and sets of objects in the domain. It does this by recursively combining basic descriptions using logical connectives and relations between objects, always keeping track of the extension sets of the logical forms constituting the descriptions. The combinations are stored in a chart to facilitate reuse of intermediate results. The chart is pruned by the requirement that every combination of properties produced has to be realisable linguistically, making this one of the rare approaches that integrate surface realisation with content selection. To guide the search for 'optimal' referring expressions, Varges suggested a number of constraints that either apply during the chart-building process or filter out unwanted solutions afterwards. Varges and van Deemter (2005) extended Varges' (2004) algorithm in order to be able to produce quantified expressions, which is, as they show, impossible in approaches such as the IA and the graph-based framework.

Description Logics (DL) constitute yet another approach to the generation of referring expressions that is based on viewing the semantics of a referring expression as a logical formula. Similar to CGs, DLs are a well-established First-Order Logic formalism; and similar to Varges' approach they have been used to combine

basic formulas (or descriptions) via logical connectives to generate descriptions for all objects and sets of objects in parallel. The first implementation of a REG algorithm in an already established DL system was mentioned in (Gardent and Striegnitz, 2007). Areces et al. (2008) worked out the different options for implementing REG in DL in more detail with a focus on the expressivity of different DLs.

Just like CGs, DLs divide the knowledge base into two parts: ontological knowledge is stored in T-Boxes equivalent to the support hierarchies in CGs, while general domain knowledge is represented in A-Boxes instead of a bipartite graph. One important difference is that DLs do not allow relations of arity higher than two. A Description Logic defines a grammar over logical formulas containing a subset of the constructs that are legal in First Order Logic. There are many different Description Logics, each with a different set of logical connectives. This set of connectives determines the expressiveness of the language. For example, only in a DL allowing negation $\neg$ would it be possible to represent a referring expression such as the one in Example 2.3, repeated here:

(2.3) the black dog that is *not* a poodle

The equivalent DL formula for this referring expression is

(2.15) $\mathsf{black} \sqcap \mathsf{dog} \sqcap \neg\mathsf{poodle}$.

Areces et al. (2008) observed that referring expressions can be found in DLs by computing what they call the **similarity set** of the target referent. The similarity set of an entity e contains all entities that have the same properties as e. Of course, if we are interested in generating a distinguishing description, we hope that the target referent is alone in its similarity set.

Areces and colleagues used a well-established efficient algorithm by Hopcroft (1971) for the computation of **simulation classes**, which coincide with similarity sets, and extended it to also generate a DL formula for each simulation class. The algorithm works by partitioning the set of domain entities into smaller and smaller subsets that each can be described using a DL formula. In each iteration it refines the formulas by adding conjunctions and disjunctions with other formulas or their negations and afterwards also relations between formulas. It terminates when no further divisions are possible, that is, when the subsets are equivalent to the simulation classes of the domain. In essence, the algorithm generates descriptions for all distinguishable entities and sets of entities in the domain at once.

Of course, there are always many different formulas that could be used to describe a certain subset. Areces et al. did not explicitly provide a mechanism

to control which description is generated beyond stating that propositions (i.e. atomic properties) should be used first, before relations are tried; however, they mentioned that it is possible to enforce a preference order over the propositions and relations, and this is what they appear to have done for the evaluation of their algorithm.

2.3 Relations in REG

Generating referring expressions that contain relations between objects has proven to be a particularly hard problem, because relations to other objects are much more complex than attributes with simple atomic values. Some of the questions complicating the task of generating relational referring expressions are: When should a relation be considered for inclusion? Which factors should the decision to include a relation be made dependent on? Which properties of the related object should be included as well?

In the context of spatial relations, the literature on spatial cognition and computer vision usually calls the object being described the **figure**, and the object that the figure is related to is the **ground**. In the REG literature, the figure usually retains the name **target referent**, and the ground is often called the **landmark**.

The task of generating referring expressions that contain relations has been addressed in a variety of ways, which I discuss in the following.

2.3.1 Relational REG Using Constraints

The first algorithm that was able to handle relations between objects was the Relational Algorithm (RA) proposed by Dale and Haddock (1991b). It is based on GREEDY from (Dale, 1989) in that it applies the same greedy heuristic and aims for the shortest possible referring expression. In order to deal with relations to other objects and the properties of these objects, the RA employs a constraint network to keep track of the distractors and a stack of objects which are to be described recursively.

The RA (see Algorithm 2.3) is initialised with a constraint network N. The properties in the description L correspond to the constraints in N. The second component of N is the set of distractor sets D_i for all objects i that are mentioned in L. When a new property p is added to N ($\mathsf{N} \leftarrow \mathsf{N} \oplus \mathsf{p}$), p is added to L and all distractor sets in N are adjusted according to the now changed constraint set. If p is a relation to another object j, a new distractor set D_j is added to N and initialised to all the distractors of j that are not ruled out by L already. The stack of objects to be described initially contains only the target referent r. The

r — the intended referent
$\mathsf{D_r} = \{\mathsf{d} \mid \mathsf{d}$ *is a distractor of* $\mathsf{r}\}$
KB — the knowledge base containing the properties of all domain objects
$\mathsf{L} = \{\}$ — the empty description
$\mathsf{N} = \langle \mathsf{L}, \{\mathsf{D_r}\}\rangle$
`stack` $= [\mathsf{r}]$ — the stack of objects to be described
$\mathsf{P} = \{\mathsf{P_r}\}$ — contains the set of properties for each object in `stack`

1. Check Success
   $\mathsf{cr} \leftarrow$ top of `stack`
   if `stack` $= \emptyset$ **then** return L
   elseif $|\mathsf{D}_{cr}| = \emptyset$ **then**
      pop `stack`
      add $\mathsf{D_{cr}}$ to N
      goto Step 1
   elseif $\mathsf{P_{cr}} = \emptyset$ **then** fail
   else goto Step 2

2. Choose Property
   for each property $\mathsf{p_i} \in \mathsf{P_{cr}}$ **do**
      $\mathsf{N_i} \leftarrow \mathsf{N} \oplus \mathsf{p_i}$
   Choose property $\mathsf{p_j}$ so that $\mathsf{N_j}$ contains the smallest set $\mathsf{D_{cr}}$
   goto Step 3

3. Extend Description (with the chosen p_j)
   $\mathsf{P_{cr}} \leftarrow \mathsf{P_{cr}} - \{\mathsf{p_j}\}$
   for every object g related to r in $\mathsf{p_j}$ **do**
      push o onto `stack`
      $\mathsf{P_o} \leftarrow \{\mathsf{p} \mid \mathsf{p(o)} \in \mathsf{KB}\}$
      $\mathsf{P} \leftarrow \mathsf{P} + \mathsf{P_o}$
   $\mathsf{N} \leftarrow \mathsf{N} \oplus \mathsf{p_j}$
   goto Step 1

Note: $\mathsf{N} \oplus \mathsf{p}$ signifies the result of adding p to the constraint network N.

Algorithm 2.3: **The Relational Algorithm (RA, Dale and Haddock, 1991b)**

algorithm always works on describing the top element of the stack, which results in a depth-first search behaviour. Whenever a relation is added to L, the new object o introduced by this relation is pushed onto the stack and becomes the top element to be described. When the top element is fully distinguished from all its distractors, it is popped off the stack and the one below becomes the current referent cr to be described.

The constraint network ensures that the distractor sets of all objects already in the description are kept up to date at all times, including the distractor set of the target referent. This often means that once a landmark is fully distinguished from its distractors, the target referent also has no distractors left. Let's look at

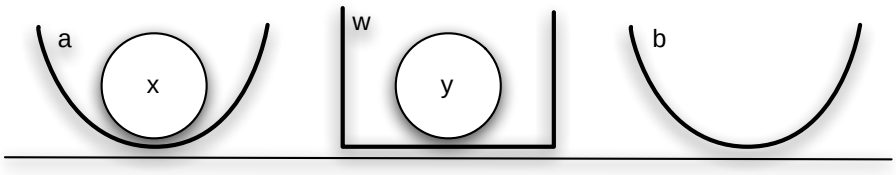

Figure 2.5: An example domain with relations

an example for the small domain in Figure 2.5 consisting of two balls x and y, two bowls a and b and a box w. Assume that we are building a referring expression for x and after two iterations of the algorithm the constraint network and the stack are:

$$\begin{array}{lcl} \mathsf{N} & = & \langle \mathsf{L} = \{\mathsf{ball(x)}, \mathsf{inside(x,a)}\},\ \{\mathsf{D_x} = \{\mathsf{y}\}, \mathsf{D_a} = \{\mathsf{w}\}\} \rangle \\ \texttt{stack} & = & [\ \mathsf{a},\ \mathsf{x}\ ] \end{array}$$

The description L asserts that the target referent x is a ball and inside another object (a); but there is another ball y which is also inside an object (w). This makes y a distractor for x and w a distractor for a. It is now object a's turn to be described. Let's assume the algorithm adds bowl(a) to L, which distinguishes a from w because w is a box. This also means that ball x cannot be confused with ball y anymore, as y is inside a box, but the description now states that the target referent is inside a bowl. At this point the constraint network and stack are:

$$\begin{array}{lcl} \mathsf{N} & = & \langle \mathsf{L} = \{\mathsf{ball(x)}, \mathsf{inside(x,a)}, \mathsf{box(a)}\},\ \{\mathsf{D_x} = \{\}, \mathsf{D_a} = \{\}\} \rangle \\ \texttt{stack} & = & [\ \mathsf{a},\ \mathsf{x}\ ] \end{array}$$

By adding a property for a, both distractor sets $\mathsf{D_a}$ and $\mathsf{D_x}$ have been cleared. In the following three iterations, Step 1 will first pop a off the stack, then x will be popped off the stack, and finally L will be returned as a distinguishing description because the stack will be empty at that point.

One detail that is omitted in the pseudocode in Algorithm 2.3 is that before a property is added to the description in Step 3, the object that this property belongs to is 'anonymised' by associating it with a variable replacing the object ID. If the property is a relation, both related objects are replaced by different variables. This ensures that it is really the properties that distinguish an object from its distractors rather than its unique ID from the knowledge base. However, in Step 2, where the discriminatory power of all candidate properties is assessed, the landmark in a relation is not replaced by a variable. Therefore, the assessment step takes into account the identity of the landmark for each relation. For example,

when deciding whether to include the inside relation for x in the above domain, the algorithm not only takes into account that x is inside another object, but also that it is inside object a. As there is no other object inside a, the inside(_,a) relation rules out all three distractors, more than if a had not been taken into account, in which case inside(_,_) would rule out only two distractors (a and w). Intuitively, it makes sense to take the landmark object into account in the decision as to whether to include a relation because this strategy implicitly takes into account the landmark's properties. However, we will see in Chapter 4 that, especially in domains in which the entities are highly connected among each other via relations, this strategy combined with the algorithm's depth-first search behaviour can lead to very unnatural referring expressions being produced.

Dale and Haddock drew particular attention to the problem of infinite regress whereby a description might run into a loop by describing an object in terms of the same relations again and again. In our example, this could be a description such as

(2.16) the ball inside the bowl containing the ball inside the bowl ...

and so forth. To avoid this happening, they proposed to prohibit any information (property or relation) from being used twice in the same referring expression. Interestingly, this seems unnecessary due to the way the constraint network keeps track of all distractor sets at once: a property or relation that is already part of the description is already accounted for in the distractor sets of all entities in the constraint network. Therefore, a relation, when considered for a second time, would not rule out any further distractors and consequently not be chosen for inclusion in Step 2 of the algorithm. All distractors the relation can possibly rule out were already excluded the first time it was added. In our example from above, when object a is added to the description, its initial distractor set only contains object w, not object b, because the description already contains the information that a contains a ball, which is not true of b. Including the relation in(x,a) again would not rule out the only remaining distractor of a (b), so the algorithm would not choose it.

Gardent and Striegnitz (2007) present a base algorithm that is a mixture of GREEDY, the IA and Dale and Haddock's (1991b) RA. This base algorithm does not specify by which criterion a property is chosen, so both the greedy and the incremental heuristic are possibilities. Their algorithm uses a recursive procedure identical to the one used by Dale and Haddock in order to be able to use binary relations between objects in the referring expressions used: a stack keeps track of

the objects that have been introduced by binary relations and for each object the basic algorithm is invoked recursively before it is popped off the stack.

2.3.2 Relational Extensions of the Incremental Algorithm

In Section 2.2.2, we already saw many extensions to the Incremental Algorithm by Dale and Reiter (1995). Most of these extensions are concerned with overcoming limitations in terms of the kind of information that the IA can include in referring expressions. I will discuss here in more detail the extensions that give the IA the capability to handle relations between objects.

Theune and Krahmer (Theune, 2000; Krahmer and Theune, 2002) were the first to present a relational extension for the IA. In their version, the algorithm gets called recursively as soon as a landmark is introduced via a relation. This produces a depth-first search behaviour equivalent to that of Dale and Haddock's algorithm. However, relations are highly dispreferred in Theune and Krahmer's approach as they chose to place spatial relations at the end of the IA's preference list.[2] This is based on what they called the 'omnipresent principle of least effort' (Zipf, 1949; Clark and Wilkes-Gibbs, 1986), from which they concluded that '[i]t seems an acceptable assumption that people prefer to describe an object in terms of simple properties, and only shift to relations when properties do not suffice', as 'it takes less effort to consider and describe only one object' (Krahmer and Theune, 2002, p. 32).

Krahmer and Theune's (2002) algorithm takes the landmark into account when deciding whether to include a relation, just as Dale and Haddock's RA does. However, it does not anonymise either of the two related objects once the relation is included. This means that in the computation of whether the landmark has been fully described the identity of the original referent is taken into account. This is different from Dale and Haddock's approach, where only the properties already included in the description are taken into account. Theune and Krahmer's approach would presumably result in few properties being included for the landmark: the landmark only has to be distinguished from objects that stand in the same relation to the target referent as the landmark, and it is unlikely that there are many of these.

The problem of infinite recursion was tackled in a similar way to the use of constraint networks: the initial distractor set for a landmark to be described takes into account the description already produced so far. Therefore, a relation already

[2] In (Theune, 2000) this is done by using separate preference lists for properties and relations and only moving to the list of relations once all properties have been tried.

included in the description would not rule out further distractors of the landmark and will not be chosen.

The lexicalised context-sensitive version of the IA by Siddharthan and Copestake (2004) also includes a recursive step for relational attributes. Just as for other attribtues, a relation's position in the preference order is determined by a discriminating quotient, this time taking into account not only the relation itself but also the related object: the fewer other objects that have the same relation to the same landmark, the higher the rank of the relation. This is the only approach to extending the IA to relational descriptions that does not resort to a simplistic solution for sorting relations into the preference ordering.

Once a landmark has been introduced into the referring expression, Siddharthan and Copestake's algorithm gets called recursively for this object. They employ a similar strategy for describing the relatum as Theune and Krahmer in that they only allow as distractors other objects that stand in the same relation to the target referent, rather than, for example, all other objects of the same type as the landmark. They point out that this strategy is helpful in domains with few objects of the same type, but in domains with a lot of objects of the same type as the landmark it might make sense to build a description that helps locate the landmark by distinguishing it from all of these objects rather than just those that have the same relation to the target referent. To avoid infinite regress the algorithm keeps track of which entities have already been used in the referring expression under construction and does not allow more relations to be included to the same objects. This strategy of excluding repeated use of the same objects was already used by Davey (1978), but as Dale and Haddock (1991b) pointed out, this makes it impossible to produce felicitous descriptions with larger loops such as the one in Example 2.17.

(2.17) the man who ate the cake which poisoned him

Kelleher and Kruijff (2005, 2006) proposed an approach to extend the IA that was based on the need for a REG module in a spatially aware autonomous agent. They cited Clark and Wilkes-Gibbs' (1986) Principle of Minimal Cooperative Effort and Dale and Reiter's (1995) Principle of Sensitivity, as well as a production study by van der Sluis and Krahmer (2004b), to motivate their proposed preference ordering. Similarly to Krahmer and Theune (2002), they argued that from these principles it follows that the cognitive load imposed by producing a relational description is higher than that imposed by processing a simpler, non-relational, description. Accordingly, their system only includes spatial (and, hence, relational)

information in a referring expression if it is not possible to construct a description from non-relational attributes. They do this by separating properties and relations into two different preference lists. The list of relations only gets considered once the list of non-relational properties has been exhausted.

Their algorithm avoids the problem of infinite regress by focussing on the choice of appropriate landmarks: when a landmark lm is being used to describe the target referent r and another relation has to be used in the recursive step to describe lm, the original target entity r is not included in the set of the potential landmarks from which the algorithm chooses one to relate to lm. Kelleher and Kruijff did not explicitly exclude the original target from being used as a landmark; rather, this behaviour is a side effect of the way they divide the set of all entities into those that can and those that cannot be used as landmarks.

Another important difference to Theune and Krahmer's approach lies in the fact that Kelleher and Kruijff took the visual and linguistic salience of the landmark into account in the decision of whether to include a relation. In its first version (Kelleher and Kruijff, 2005), their algorithm iterates through the preference order of relations and for each relation considers all landmarks that have this relation to the target in order of the salience of the landmarks. In (Kelleher and Kruijff, 2006), more emphasis is put on the salience of the landmark by iterating first through the list of landmarks (ordered by salience) and then considering the relations that each landmark has to the target referent in order of preference. This strategy makes it possible to mimic Theune and Krahmer's subsumption hierarchy over spatial relations but is not limited to it, as it allows ordering according to other criteria than semantic subsumption; Kelleher and Kruijff suggested that the relations should be ordered by the cognitive load they impose on speaker and listener. For example, psycholinguistic evidence shows that relations in the vertical dimension are preferred over horizontal ones (c.f., Lyons, 1977; Bryant et al., 1992; Gapp, 1995; Bryant et al., 2000; Landau, 2003; Arts, 2004; Tenbrink, 2004).

These relational extensions of the IA all have one problem in common which has been pointed out before by Krahmer and Theune (2002), Krahmer et al. (2003) and van der Sluis (2005): they extend the concept of incremental processing from atomic one-place attributes to relations. The lack of a backtracking mechanism means that any relation that excluded at least one distractor at the time it was considered for inclusion will appear in the final referring expression, even if it is rendered redundant by further relations included later. While incremental content selection might intuitively be an appropriate behaviour for non-relational attribtues, it can easily produce relational descriptions that sound much less plau-

sible. For example, these extensions might generate a description such as Example (2.18) when (2.19) might have sufficed. In (2.18) *next to the bush* becomes redundant once *under the tree* was included, but the algorithm cannot exclude it retrospectively.

(2.18) the dog next to the bush under the tree

(2.19) the dog under the tree

2.3.3 Relations in Graphs

Approaches to REG that represent context sets as graphs, such as the two discussed in Section 2.2.3, are particularly well-suited for the generation of referring expressions that contain spatial relations. Their main advantage is that they do not run into the problem of infinite recursion as many algorithms using a propositional knowledge representation do: the graph-based frameworks do not require a recursive call of the base algorithm, and an edge representing a relation between two objects can by definition only ever be included once in a subgraph.

In Krahmer et al.'s (2003) framework, binary relations are represented as edges between the nodes representing the related objects. Non-relational properties are represented as loops, edges that originate and end in the same node. This means that no special mechanism is needed to deal with relations; they can be treated in the same way as other properties.

Whether a relation is included in the referring expression under construction depends on the cost of the relation combined with that of the properties of the landmark relative to the costs of other properties of the target referent. van der Sluis and Krahmer (van der Sluis and Krahmer, 2005; van der Sluis, 2005) suggest making relations more expensive than other properties for a similar reason to that underlying the relational extensions of the IA discussed in Section 2.3.2, which place relations at the end of the preference order: they assume that inherent properties are easier to perceive and process than relational ones.

Croitoru and van Deemter (2007) make no mention of a mechanism to influence the order in which properties are included in the referring expression or which referring expression should be preferred in their CG approach, but it seems to be straightforward to use the greedy or incremental search algorithms on the CG representation or to add a cost function similar to the one used by Krahmer and colleagues.

In terms of dealing with relations, the main advantage of the CG approach is that it is the only framework that naturally allows for relations of higher arity than two. So far, we have only considered spatial relations between two objects at a time. The standard example of a relation involving more than two objects is *between*. In Example (2.20) a spatial relationship is described between three entities: a dog, a tree and a bush. If we move away from visual scenes and spatial arrangements of objects, relations of even higher arity might need to be accounted for. For example, (2.21) involves four entities: a treaty and the three countries who have signed it. This example could of course be extended to include any number of countries, so the arity of the relation is not predefined.

(2.20) the dog between the tree and the bush

(2.21) the treaty between France, Germany and the UK

2.3.4 Relations in Other Approaches

In Varges' (2004) overgenerate-and-rank approach, relations are treated similarly to logical connectives that can combine two logic formulas into a more complex description. In this approach all possible descriptions for all objects and sets of objects are generated in parallel. Varges defined a combination rule that takes a pair of descriptions d_1 and d_2 and a relation *rel* and establishes whether there are any objects in the extensions of the two descriptions which can also be described by a relational description derived by combining the original two descriptions using a relation. For example, let $d_1 = \langle \mathsf{bowl} \rangle$, $d_2 = \langle \mathsf{table} \rangle$, $rel = \mathsf{on}$, and the extensions of d_1 and d_2 be $ext_{d_1} = \{\mathsf{b_1}, \mathsf{b_2}\}$ and $ext_{d_2} = \{\mathsf{t_1}, \mathsf{t_3}\}$, respectively. Now let us assume that $\mathsf{b_2}$ is on top of $\mathsf{t_1}$. That means that d_1 and d_2 can be combined to $d_3 = \langle \mathsf{on}(d_1,d_2) \rangle = \langle \mathsf{on}(\langle \mathsf{bowl} \rangle, \langle \mathsf{table} \rangle) \rangle$, which has the extension set $ext_{\mathsf{d_3}} = \{\mathsf{b_2}\}$.

Varges proposed a monotonic increasing cost function based on the number of words in the surface form of a referring expression to steer the generation process away from duplicated properties, in a similar way to that in which the cost function in Krahmer et al.'s (2003) graph-based framework guides the search. This cost function can then also be used to avoid the problem of potential infinite regress that is introduced by the use of relations: the monotonicity constraint on the cost function ensures that a description containing the same relation twice is always more expensive than one containing the relation only once. It is to be expected that this cost function will prefer non-relational properties over relations, as a relation usually introduces at least two words, while many non-relational properties can be expressed as a single adjective.

In (Varges, 2005) the overgenerate-and-rank approach is applied to the problem of describing points on MapTask (Anderson et al., 1991) maps at which the path takes a turn. As these points do not have any other properties, the only way to describe them is in terms of their spatial relations to the landmarks on the map. For the purpose of this exercise the maps are overlaid with a grid that effectively 'pixelates' them. Each landmark and each target point covers a set of pixels. The aim is to describe the set of pixels that make up the target point. An example description in this domain would be *[the points] above the west lake and to the left of the great viewpoint.* Two non-empirical criteria are proposed for ranking: the ratio of the extension size of a description to the number of pixels in the target point; and the number of characters in the surface form, which results in a search for shorter, non-redundant, descriptions. The author points to the possibility of using the MapTask corpus as a source of more empirically motivated ranking criteria in future work.

Binary relations are a natural component of the Description Logics $\mathcal{EL}$ and $\mathcal{ALC}$ for which Areces et al. (2008) presented REG algorithms. These algorithms are in no danger of running into infinite regress as they do not involve recursion or backtracking. Areces et al. suggested first dividing the domain objects into sets by building descriptions from atomic properties only and then adding relations if necessary. This mimics the same dispreference for relations incorporated in most of the relational extensions of the IA as well as the cost functions that have been proposed for the graph-based framework and Varges' (2004) overgenerate-and-rank approach.

2.4 The Use of Corpora in REG

Most of the 'classic' approaches to content selection in REG were developed based mainly on the authors' intuitions with regard to what content referring expressions should ideally contain, and loosely guided by psycholinguistic principles such as Grice's Conversational Maxims and Zipf's Principle of Least Effort. None of the basic frameworks and algorithms for content selection have sound empirical backing, and only recently have some of them been evaluated against human-produced data (Jordan and Walker, 2005; Gupta and Stent, 2005; Viethen and Dale, 2006a; Gatt et al., 2007; van der Sluis et al., 2007).

In this section, I take a look at existing human-produced corpora of referring expressions and some previous work on REG that involves these corpora either in the development of algorithms or in evaluating their output.

2.4.1 Existing Corpora

A number of corpora exist that contain human-produced referring expressions. I briefly describe those corpora that are publicly available or can be obtained from the researchers who collected them.

The MapTask corpus (Anderson et al., 1991) is a collection of 128 dialogues between an instruction giver and an instruction follower. Each dialogue partner had a map of the same environment showing a number of landmarks. Their task was to reproduce on the follower's map a path that was only shown on the giver's map. For some dyads there were a few mismatches between the maps in the form of missing landmarks and differently named landmarks. The landmarks were chosen and labelled with names based on criteria that would facilitate the study of phonological phenomena. The labelling of the landmarks makes the corpus less suitable for the study of content selection for referring expressions, as the participants can rely on the labels provided rather than being required to build distinguishing descriptions themselves. Nonetheless, Gupta and Stent (2005) annotated the referring expressions in a subset of 30 dialogues for the use of modifiers and evaluated a number of content selection algorithms on this subcorpus (see Section 2.4.3).

Di Eugenio et al. (2000) collected the COCONUT Corpus, which consists of 24 computer-mediated two-person dialogues. Each dialogue partner was given a budget and an inventory of furniture to choose from, and the task was to jointly buy furniture for the living and dining rooms of a house. In the transcripts of the resulting dialogues, each reference to a furniture item was annotated to indicate the attributes used and whether it was an initial reference or a subsequent anaphoric reference. A second annotation at the utterance level captures the state of the problem solving process and divides each dialogue into segments according to the changing purpose of the utterances. Utterances are also annotated for their discourse functions, recording the level of commitment each utterance puts on the participants for a certain buying action.

In these dialogues, speakers and listeners did not share a visual display of the described objects as was the case for the MapTask and most other corpora I discuss in this section, but rather each speaker sees a different set of objects. Furthermore, the referring expressions in this corpus fulfil other functions than simply identification. For example, in an utterance such as *let's go with the $150 table*, the price is often included to convince the listener that this action should be preferred over buying a more expensive table. The corpus is particularly well-suited

for the study of the impact that changing discourse factors such as task-specific goals and problem solving states have on human-produced referring expressions, which is what Jordan and Walker used the corpus for in their machine learning experiments described in Section 2.4.2 (Jordan, 1999, 2000a,b; Jordan and Walker, 2000, 2005). The richness of the context makes it difficult, however, to factor out the impact of prior discourse and conflicting functional aspects of reference on the task of identification.

The Bishop Corpus, described in (Gorniak and Roy, 2004), is a collection of object descriptions designed to learn a model for reference understanding. Participants were shown a computer-generated 3D rendering of a scene with up to 30 green or purple cones. They were asked to describe one of the cones which would then be removed from the scene. It contains a development set of 268 descriptions and a test set of 179 descriptions. As the cones visually only differed in their colour, all referring expressions in this corpus contain some sort of locational information. However, only 6% of descriptions in the development set mention a spatial relation to another cone. As all objects had the same visual appearance, this corpus is not very well suited for a study of the impact that the visual salience of an object has on the likelihood of it being used as a landmark in a description. A further disadvantage is that the positions of the cones were determined randomly and that the participants could choose freely which cone they wanted to describe. This means that the corpus is not balanced for the positions of the target cones within the scenes. Furthermore, the scene did not change for each trial, creating a temporal dependence effect between the descriptions. For example, Gorniak and Roy (2004) mention that 4% of the descriptions contained a reference to the previous target referent, which had already been removed from the scene.

The first corpus that was specifically designed for the study of the semantic content of referring expressions with identification as their sole function was the TUNA Corpus (van Deemter et al., 2006; van der Sluis et al., 2006). For this corpus, a controlled language production experiment was conducted, similar to those I describe in Chapter 5. Here each referring expression was collected in isolation rather than as part of an extended discourse. Participants saw a 2D display showing a number of entities and were asked to type a description either for a single entity or a pair of two entities, which were highlighted on the screen. To make the task believable the participants were told that the experiment was testing an automatic text understanding system. Two different domains were used for this corpus: in the furniture domain, the stimuli showed common furniture

items which could be distinguished by their type, size, colour, and the direction they were facing; the people domain showed black and white photographs of men who could be distinguished by the clothing items they were wearing and hair and beard colour (dark or light). The corpus contains 780 descriptions of singular target referents (420 in the furniture domain and 360 in the people domain) and 1500 plural descriptions (780 and 720 for furniture and people domains, respectively).

The main design factor was fault-criticalness (FC), where two thirds of the participants had only one shot at describing the target referents (+FC), while the other third were given the opportunity to repair descriptions when the system did not pick out the correct item (–FC). Half of the participants in the +FC condition were told that the system had access to the same domain layout as they did, which made the location of the target referents a distinguishing feature (+LOC). All other participants were told that they could not use location in their descriptions (–LOC). In half of the 1500 plural trials the two targets were similar in that the shortest possible non-locational description was the same for both. In the other half, the two targets were more dissimilar.

Each trial was annotated with full knowledge base information for all entities in the domain as well as the semantic content of the collected referring expression. This makes the corpus a very useful resource for automatic evaluation of REG systems. The singular part was used for the three REG evaluation challenges in 2007, 2008 and 2009 (see Sections 2.4.3 and 4.4), while the plural part of the corpus has been used to study the way people group multiple target referents in referring expressions (Gatt, 2007). Due to the TUNA design, the singular trials of the corpus lend themselves to the study of the impact of fault-criticalness and discriminatory power of properties on the construction of non-situated referring expressions. However, the relatively small number of trials in which location was a useful feature (260 singular trials) makes it hard to draw conclusions about the use of location, and the layout of the stimuli successfully discouraged the use of spatial relations between entities.

A second corpus that has been used in the REG evaluation challenges is the GREC Corpus (Belz and Varges, 2007a,b). It is a collection of introductory texts from Wikipedia about a variety of types of entities, including people, rivers, cities, countries, mountains and lakes. Each reference to the main subject of each text is annotated with its syntactic case and the form of referring expression (name, common noun, pronoun, or empty). This corpus is aimed at the planning task that has to take place before content selection for a distinguishing referring expression might begin: choosing whether a full-fledged noun-phrase is indeed necessary or

whether a name or pronoun can be used. Its focus on encyclopaedic texts also means that the target referents always have a name in the form of the text's title, which makes it not ideal for the study of content selection for distinguishing descriptions.

The iMap corpus, which has been mentioned in a number of publications (Guhe and Bard, 2007; Louwerse et al., 2007; Guhe and Bard, 2008a,b; Viethen et al., 2010), is an adaptation of the original MapTask. Here the maps were more densely populated with unlabelled landmarks and the paths were longer and more complex, which resulted in more landmark mentions. The landmarks were designed in a way that encouraged the speakers to choose from a set of predefined properties to distinguish them from each other, which makes the corpus interesting for the study of content selection. The corpus consists of 256 dialogues with 26,488 singular referring expressions and 18,726 plurals, which are annotated with their referent IDs and the properties used.

The format of the corpus lends itself to the study of patterns that evolve over the course of a dialogue or even a number of dialogues, such as tracking the use of individual properties in initial references as influenced by different discourse factors (Guhe and Bard, 2008b,a) or modelling alignment processes within coreference chains (Viethen et al., 2010). This corpus is not yet publicly available, but the original authors have mentioned plans to publish it.

2.4.2 Empirical Approaches to REG

The first corpus-based approach to referring REG was presented by Passonneau (1995, 1996), who analysed the referring expressions found in a corpus of short spoken narratives to inform a model for the generation of referring expressions. Her model was based on Dale's (1989; 1992) algorithm for generating minimal distinguishing descriptions, but augmented it to take into account focus-structure information Grosz et al.'s (1983) Centering model, in order to achieve an improvement in the choice of the form of reference that the model generated. She tested her integrated model by assessing its accuracy in reproducing the distribution of pronouns, minimal descriptions and over-specified descriptions against that found among 319 subsequentreferring expressions from a held out set of three narratives. She found that her integrated system was better able to replicate the forms of references found in the test set than two baseline models that did not have access to the full focus structure derived by the Centering model.

Jordan and Walker (2000, 2005) presented the first machine learning approach to content selection for REG. They used the rich discourse-sensitive annotation of the COCONUT Corpus to define features for the machine learning system RIPPER (Cohen, 1996) which automatically induces rules from data observations. Three types of features were used: contrast set factors, which record which objects are currently in the distractor set according to Grosz and Sidner's (1986) model of discourse structure; conceptual pact factors, which are inspired by the lexical alignment model of Clark and colleagues (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996); and intentional influences factors, which are based on a model developed by Jordan (2000a).

Jordan and Walker used 25-fold cross-validation on 393 referring expressions from 13 of the COCONUT dialogues to test different combinations of features. They measured the absolute accuracy, this being the proportion of referring expressions generated that are identical to the human-produced reference descriptions from the corpus. In isolation, the intentional influences factors performed better (42.4% accuracy) than the other two feature sets (contrast sets: 30.4% and conceptual pacts: 28.9%) and combining the three types of features did not significantly increase accuracy (43.2%). However, what had the highest impact was a fourth, theory-independent, type of features that recorded trial-specific information, such as the trial ID, the participant-dyad, the actual speaker and the exact attribute values of the target referent. In isolation, this collection of features achieved 54.5% accuracy, and combining them with all three other types of features only increased this performance to 59.9%. These results lend mild support to Jordan's intentional influences model over the other two models, but most strongly suggest that none of the models capture the variation in the data very well.

Another foray into using machine learning for REG was made by Stoia et al. (2006). They aimed at building a dialogue system for a situated agent giving instructions in a virtual 3D world. However, this approach was not focussed so much on content selection as on determining the best form of reference to use. They used a machine learner to train decision trees that decided which determiner to use, what type of head to include in the noun phrase (e.g. a pronoun or a common noun) and whether or not to use a modified noun phrase. The semantic content of the modifier was not at issue. The features available to the decision tree learners were a mix of dialogue history, visual context and semantic type information about the target referent. They trained separate decision trees for determiner, head noun and modifier choice and applied them sequentially, with each

tree having access to the output of the previous tree. For training and automatic evaluation they used a set of 1242 referring expressions from a collection of dialogues between two conversation partners who were carrying out the instruction task in the same virtual world as the system would be employed in later. This automatic evaluation found that the decision trees were able to match the human data in 31% of all cases. As they were interested not so much in the human-likeness of their system, but mostly in its effectiveness, they also performed an intrinsic human evaluation in which participants were asked to compare the system output to the human-produced referring expressions and a random baseline. The human evaluators judged 62.6% of the referring expressions generated by the system to be as good or better than the human-produced references.

A number of the systems submitted to the REG evaluation challenges based on the TUNA Corpus were based on empirical analyses of the training set. Most of these systems were based on the IA and used a simple frequency count of the properties in the training set to inform the order in which the target referent's properties should be tried (Kelleher, 2007; Spanger et al., 2007; Fabbrizio et al., 2008; Kelleher and Namee, 2008; de Lucena and Paraboni, 2008; Gervás et al., 2008; de Lucena and Paraboni, 2009). One team, which I was part of, used frequency-based cost functions in the Graph-Based Algorithm (Theune et al., 2007; Krahmer et al., 2008; Brugman et al., 2009). Bohnet (2007, 2008, 2009) combined nearest-neighbour learning with a full brevity approach, in order to pick the shortest referring expression that best matches the training data for a given a target; and used a decision-tree learned from the training data to dynamically determine the preference order for the IA. In 2008 and 2009, Bohnet tailored his full brevity algorithm to match individual participants, but found that participant information was not reliably provided in the test data. Fabbrizio et al. (2008) presented the only other approach which attempted to capture speaker-specific preferences in the full brevity and the incremental algorithm. Their full brevity approach picked the shortest descriptions that was either most often or most recently used by the same speaker, and their version of the IA used speaker-specific frequency-based preference orders. King (2008) and Hervás and Gervás (2009) used evolutionary programming for the content selection task, but both were met with very limited success.

Because the proceedings only allowed 2–4 pages for the non-peer reviewed description of each of these systems, not much detail is available for most of the approaches. Many of them appear to use rather similar techniques and are based on the same traditional algorithms. It would be interesting to see more comprehensive accounts of how they work and how exactly they differ from each other.

2.4.3 Evaluation against Human-produced Data

As mentioned before, traditionally the output of REG systems has not been evaluated against corpora of human-produced referring expressions. Only recently has a trend towards corpus-based evaluation started to emerge in the form of a short series of evaluation competitions and a small number of papers in which the output of existing algorithms was compared to corpus data.

Gupta and Stent (2005) evaluated the IA, a greedy version of the IA based on Siddharthan and Copestake's approach of ordering the preference order by discriminatory power (SC), and a number of variants of the two on a set of automatically extracted and hand-annotated noun phrases from the MapTask and the COCONUT dialogues. They compared the algorithms to a baseline that included the type of the target referent and then randomly picked properties until the description ruled out all distractors. The variants they introduce are based on the observation that partners adapt their referring expressions to each other (Brennan, 1996; Metzing and Brennan, 2003): in one version they re-ordered the preference list to the order in which the properties appeared in the previous mention of the target referent, and in the second version they additionally forced the algorithms to reuse all properties that were used in the previous mention.

A second type of modification addressed one aspect of surface realisation by coupling the content selection components of the IA and SC with pre- and post-modifier ordering. This aspect was also taken into account in the evaluation metric they used, which computed

$$S = \frac{C}{C + I + D + M} \tag{2.22}$$

where

C = the number of correct attributes
I = the number of inserted attributes
D = the number of deleted attributes
M = the number of moved attributes.

They found that their modified versions of the algorithms, which adjusted the modifier ordering, performed better than the basic versions of IA and SC. This is not surprising as neither IA nor SC were designed to address the surface ordering of the attributes they include. Clearly, using an evaluation metric that takes ordering into account, here by including the factor M, disadvantages approaches that are 'purely semantic' (Gatt, 2007, p. 97). A second concern I share with Gatt (2007)

is the fact that Gupta and Stent only test the IA with one preference order for each data set and do not motivate their choice of this particular order. As mentioned in Section 2.2.2, the choice of preference order can have a large impact on the content of a referring expression produced by the IA.

Gupta and Stent also found that their type+random baseline system outperformed the IA and SC on the MapTask data. This confirms my concerns with the suitability of the MapTask corpus for content selection tasks: because the landmarks on the maps were labelled, simply using the label (or type) of the target referent usually matches the human data.

On both data sets the basic algorithms were outperformed by the versions that took partner-specific adaptation into account. This shows that the IA is too simplistic a model for the generation of subsequent referring expressions in dialogue; more sophisticated models are necessary to account for the co-operation and alignment processes that take place in such settings (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996; Pickering and Garrod, 2004). This result demonstrates that data sets of referring expressions extracted from dialogic contexts are not well suited for the study of the more basic problem of content selection in one-off or initial identification tasks.

More recently, Gatt et al. (2007) used 900 instances from the furniture domain of the TUNA Corpus to evaluate three of the classic REG algorithms against human-produced data. Their evaluation study is similar to the one that I present in Chapter 4. The three algorithms they tested were the Full Brevity algorithm (FB) that Dale (1989) aimed for, the GREEDY heuristic that Dale (1989) proposed to approximate the shortest possible description, and the IA (Dale and Reiter, 1995). To test the IA, Gatt et al. devised a number of preference orders which were based on psycholinguistic findings to the effect that colour is often used even if it is not necessary (Pechmann, 1989; Eikmeyer and Ahlsén, 1996) and relative properties such as size are often omitted (Belke and Meyer, 2002).

To compare the property sets produced by the algorithms to the instances from the corpus, the authors used the DICE coefficient, a set-comparison metric that delivers values ranging from 0 to 1.[3] They report the mean and mode (most frequent) DICE scores as well as the perfect recall percentage (the proportion of perfect matches between system output and human referring expressions). As a baseline system they used the IA choosing distinguishing properties at random rather than working off a preference order.

Overall, the results indicate that the IA is a more likely model for the process

[3]For more details on DICE and other evaluation metrics used in REG, see Section 3.3.

that humans apply to describe objects than FB or GREEDY, and they lend support to the psycholinguistically motivated choice of preference orders by Gatt and colleagues. However, they also show that, even with the best-performing preference order tried, the IA is still very far off from fully replicating the human data. While the authors showed that their preference orders beat their baselines, there is no way of telling whether another, untested, preference order might exist that would perform even better. However, it seems unlikely that the IA would be able to replicate the complete TUNA Corpus using just one preference order. As Gatt et al. (2007) point out, 'the success of the [IA] depends crucially on a balancing of ingredients that differs from case to case' (p. 56), so it might be necessary to adjust the preference order for each participant or even each trial if the IA is to be taken as a model of human reference production. Without trying all possible preference orders, it is impossible to tell from this experiment whether the IA is capable *in principle* of replicating all human referring expressions in a given corpus. I will return to this question in Chapter 4.

A second concern arises from the fact that this evaluation was carried out by first separating the data set into those descriptions that included information about the location of the target referents (412 instances) and those that did not (478 instances) and then testing the algorithms on the two different sets. Only when tested on the locational data set were they given locational attributes to choose from. This means, in effect, that the algorithms were given an unfair advantage by being told in advance whether the target description was to contain location or not. As long as all of them are given the same advantage, this might not be a big problem for comparing the algorithms; but it does distort the overall results with respect to the question of how well the algorithms can replicate human-produced referring expressions.

It is also unclear how Gatt et al. arbitrate between multiple shortest solutions of the same length in FB and GREEDY, just as Dale (1989) did not mention what his algorithm should do in this case. This plays no role in the non-locational part of the corpus, because TUNA was designed so that each target referent would have exactly one shortest non-locational description. If location is allowed, however, there might well be two competing shortest descriptions, such as

(2.23) the leftmost chair {⟨location:left⟩,⟨type:chair⟩}

(2.24) the green chair {⟨colour:green⟩,⟨type:chair⟩}

Presumably, a random choice was implemented or an implicit preference order over the properties influenced which solution was found first and returned. In either case, the performance of the algorithms compared to the human data is likely to

be impacted by the decision taken, which would make it interesting to know how it was reached. Again, I will return to this issue in Chapter 4.

The most extensive evaluation exercises on content determination for REG that have been carried out to date were a series of three public shared-task evaluation challenges. Their setup is described in (Belz and Gatt, 2007; Gatt et al., 2008, 2009a). In similar fashion to evaluation campaigns in other fields, researchers were encouraged to submit their systems to be evaluated against a subset of the TUNA Corpus. The first challenge in 2007[4] focussed solely on attribute selection; the second one in 2008[5] allowed participants to choose whether to submit systems for attribute selection only, realisation only, or end-to-end REG combining the two tasks; and the 2009 challenge[6] invited only end-to-end systems.

The attribute (or content) selection systems were evaluated for human-likeness using the set comparison metric DICE and in 2008 also MASI.[7] In all three challenges, systems were also evaluated extrinsically for identification accuracy and speed in experiments where human participants had to identify the target referents based on descriptions produced by candidate systems. In 2007, the output of the content determination systems was realised linguistically using a standardised realiser for the purpose of this extrinsic evaluation. Additionally in 2008 and 2009, Levenshtein distance, BLEU and NIST were used to compare the output of end-to-end systems to the human reference descriptions. The 2009 evaluation also involved an intrinsic evaluation by human participants who had to judge the clarity and fluency of the systems' output.

2.5 Optimality of Referring Expressions

In the majority of cases, a given target referent can be distinguished from the distractors in the context set by many different referring expressions. Which of these is returned by an algorithm depends firstly on the capabilities of the algorithm to deal with certain types of attributes and secondly on the definition of **optimality** that was assumed when the algorithm was developed. The underlying understanding of what makes a referring expression optimal determines how an algorithm makes a choice between different referring expressions that identify the target referent. In most cases the definition of optimality shapes every step of an algorithm as it chooses between preliminary descriptions that are co-extensive but

[4] `http://www.csd.abdn.ac.uk/research/evaluation/`
[5] `http://www.nltg.brighton.ac.uk/research/reg08/`
[6] `http://www.nltg.brighton.ac.uk/research/genchal09/tuna/`
[7] See section 3.3 for details on all evaluation metrics mentioned here.

do not yet exclude all distractors. Gatt (2007, p. 28) describes how the definition of **adequacy** underlying an algorithm manifests itself in an ordering over all possible distinguishing descriptions, with the optimal description to be chosen at the top of the ordering. This ordering of descriptions can, in turn, be transformed into a preference ordering or cost function over the properties which guides the algorithm's search over the space of properties.

In the following I discuss three possible interpretations of what constitutes an optimal distinguishing description. The first one is the view, first proposed by Dale (1989) and adopted by many following approaches, that the shortest description is the optimal one. The second interpretation takes a listener-oriented perspective whereby REG algorithms should aim at catering to the needs of the listener, while the third is concerned with modelling what people actually do and therefore takes a speaker-oriented or cognitive perspective. These latter two interpretations are much less clearly defined than Dale's minimality approach, but as we will see in Section 3.2.2, minimality is not as clear-cut a concept as one might think either.

2.5.1 Aiming for Brevity

Following Grice's second Maxim of Quantity (Grice, 1975), which reads 'do not make your contribution more informative than is required' (p. 65), Dale (1989) aimed at producing the shortest possible distinguishing description in every case. As we saw in Section 2.2.1, he proposed an algorithm that approximates this goal by choosing properties sequentially based on their discriminatory power. In each iteration, the algorithm includes the property that rules out most of the distractors to ensure that all distractors are ruled out as fast as possible; but it can happen that a property chosen at the beginning is made redundant by the combination of properties included after it in which case the algorithm fails to meet its brevity constraint (c.p. Reiter, 1990a; Dale and Reiter, 1995).

Based on the discussion in (Dale and Reiter, 1995), Gatt (2007) presented an algorithm that really achieves Full Brevity, but suffers from exponential worst-time complexity. Instead of constructing a referring expression by adding properties one at a time, it first checks if any single-property description suffices, then checks all combinations of two properties, then all descriptions containing three properties and so on. Effectively, the search space of this algorithm is that of all possible distinguishing descriptions for the target referent, of which there are $2^n - 1$ for n properties, rather than just the space of properties.

Reiter (1990a,b) suggested an algorithm, dubbed **Local Brevity**, which takes a distinguishing description and optimises it for length. It can take the output

of GREEDY as a starting point and then removes any redundant properties and replaces combinations of properties for single properties, if this is possible. Local Brevity guarantees a shortest possible description while avoiding the complexity issues that Full Brevity suffers from.

Based on psycholinguistic evidence (Ford and Olson, 1975; Whitehurst, 1976; Sonnenschein, 1985; Pechmann, 1989; Levelt, 1989), Dale and Reiter (1995) abandoned brevity as the main aim of referring expression generation in favour of an incremental model that in some cases produces over-specified descriptions. Although this model has become the standard model in REG, it and its extensions have been criticised for generating excessively complex, unnatural sounding descriptions in situations where simpler ones could be found. Both Gardent (2002) and Horacek (2003, 2004) provided examples where this is the case, in particular for plural target referents. Gardent came to the conclusion that it is preferable to return to the aim of minimality and implemented a constraint-based algorithm to achieve it. While Horacek did not abandon the incremental approach altogether but rather built on it, he introduced new constraints that favour shorter descriptions over longer ones.

Although there is some limited support for the aim of maximal brevity in referring expressions from experimental work in psycholinguistics (Olson, 1970, 1972; Sonnenschein, 1982), the bulk of the evidence suggests that it is a bad model of what older children and adult speakers actually do (Ford and Olson, 1975; Levelt, 1989; Pechmann, 1984; Schriefers and Pechmann, 1988; Pechmann, 1989; Arts, 2004; Maes et al., 2004; Engelhardt et al., 2006) and that it does not always produce results that are optimal for listeners (Sonnenschein, 1982; Sonnenschein and Whitehurst, 1982; Sonnenschein, 1984; Engelhardt et al., 2006). It is certainly not the case that the human descriptions found in the TUNA corpus and the corpora I will discuss in this thesis are always as short as possible.

Perhaps the best way to think of Grice's Maxims is as rules for an idealised form of communication rather than an exact description of what people actually do. In many cases, apparent discrepancies between the maxims and the behaviour of people and algorithms are mostly due to too literal an interpretation of the maxims which does not take into account all the possible situational factors. For example, when a referring expression is more helpful for a listener because it includes a redundant yet visually very salient property, it does not violate the second Maxim of Quantity ('Do not make your contribution more informative than is required'), if we interpret 'not more information than is required' to mean 'required for the listener to easily identify the target referent'.

2.5.2 Taking into Account the Needs of the Listener

Application-focussed approaches are the ones most likely to put the needs of the listener (or user of the application) at the centre of their definition of optimality. Much work in the wider field of Natural Language Generation, which is aimed at developing NLG systems that aid humans in their everyday tasks, falls into this category (see, for example Reiter et al., 2003; Yu et al., 2004; Hardcastle and Scott, 2008; Gatt et al., 2009b). Reiter and Sripada and colleagues, in particular, have discussed the impact that between-speaker variation has on the development and evaluation of NLG applications aimed at end-users (Reiter and Sripada, 2002b,a; Reiter et al., 2005). Within REG the needs of the listener is sometimes addressed in evaluation efforts involving humans. However, only very few content selection approaches to REG make an explicit attempt at addressing the preferences that a listener might have for one referring expression over another.

Dale (1990, 1992) introduced a **principle of sensitivity** which stated that the chosen referring expression should take into account the state of the hearer's knowledge. This principle was implemented by Dale and Reiter (1995) in their UserKnows function, which queries a hearer model at run-time to ensure that the hearer knows about or is able to perceive each attribute value before it is included in the referring expression under construction.

In a way, this UserKnows function can be considered a necessary requirement for any REG algorithm. In fact, most approaches take it as a given that the listener shares the same knowledge base as the speaker and therefore has access to the same information about the domain. If the listener did not know and had no way of finding out that the target referent possesses the properties mentioned in a referring expression, he would be doomed to fail at the identification task. However, one statement in (Dale and Reiter, 1995, p. 6) seems to go beyond this minimal requirement: they state that object identification 'should not require a large perceptual or cognitive effort on the hearer's part'.

This statement might be understood to hint at a need for a preference for properties that are particularly visually salient over less salient ones, although Dale and Reiter stopped short of introducing such a requirement in the IA. However, if the aim is to aid the listener in his task of identifying the target referent in a visual domain, using the most visually salient properties of the referent seems like a self-evident strategy. In large domains in particular, mentioning features that make the referent stand out from the other objects would be more helpful to a listener than a more obscure description, even if it is shorter than a visually salient one.

Interestingly, at around the same time as Dale and Reiter first published the IA (Reiter and Dale, 1992; Dale and Reiter, 1995), Edmonds (1993, 1994) proposed a plan-based model for collaboration on referring expressions that formalised the requirement of using visually salient properties. In his model each participant in a dialogue attaches confidence values to the properties of the target referent which are determined by their 'visual prominence' (Edmonds, 1994, p. 1120) and represent the participant's confidence that the property contributes to the identification of the target referent. The confidence values of the properties contained in a referring expression are combined to determine the confidence value for the expression as a whole. Only if a speaker's confidence in the whole referring expression meets a certain threshold does this speaker deem the description to be adequate. This approach favours the inclusion of visually salient, or prominent, properties in order to make identification as easy as possible for the listener. Edmonds implements a fairly simple concept of salience by defining a salience hierarchy for the properties of each type of object. This is reminiscent of Dale and Reiter's (1995) preference list of properties with the important difference that in the IA only one preference list is defined for a whole domain, rather than individual ones for different object types.

A second instance of the incorporation of the visual salience of properties in a REG algorithm can be found in van der Sluis' (2005) approach to multimodal reference. She adopted Beun and Cremers's (1998) concept of **inherent salience** for visually available objects: an object is inherently salient if it has a value for one attribute that is different from all other objects' values for that attribute. She combines this with the **focus space salience** of the target referent, which is defined by its proximity to the focus of attention (in her model the last mentioned object), and its **linguistic salience** (Krahmer and Theune, 1998, 2002; Theune, 2000). However, van der Sluis does not base the choice of individual properties on their visual salience, but rather uses the overall salience of an object to determine whether it would be useful as a landmark in a relational description as well as to decide whether it might be referred to by a pronoun or by a reduced set of properties. Clearly, this approach assumes reference in a discourse setting as it requires knowledge about previous mentions to determine focus space salience and linguistic salience.

Kelleher and Kruijff (2006) adopt a similar mix of discourse and visual salience to that used by van der Sluis and integrate it into their version of the IA intended for use in a conversational robot. They modify the stop condition for the algorithm to take into account the relative salience of the target referent in the same way as Krahmer and Theune's approach (Krahmer and Theune, 1998, 2002; The-

une, 2000) does (cf. Section 2.2.2). Also similarly to van der Sluis' approach, Kelleher and Kruijff use object salience to determine useful landmarks. A second mechanism through which Kelleher and Kruijff hope to reduce the cognitive burden on the listener is the preference ordering over properties. They base their ordering on psychological findings about the perceptual and cognitive load that different kinds of attributes (cf. Section 3.2.1) and different spatial relations place on interlocutors.

An arguably more empirical, yet also more indirect, method to ensure that an algorithm caters to the needs of the listener in REG is to let human participants evaluate its output. This can be done intrinsically by asking people to judge the referring expressions for their effectiveness, but if listener needs are at the focus of the evaluation it might be advisable to rely on extrinsic measures such as the time it takes participants to identify the correct referent and the accuracy with which they are able to identify it given a certain referring expression. Human evaluation is used extensively in other areas of NLG (one of the most extensive examples is the GIVE Challenge on instruction giving in a virtual world; Byron et al., 2009), and was the most common form of evaluation in REG until the inception of the REG evaluation challenges (Belz and Gatt, 2007; Belz et al., 2008; Belz and Gatt, 2008; Gatt et al., 2008; Belz et al., 2009; Gatt et al., 2009a).

One rare approach to content selection in REG which was both developed and evaluated with the needs of the listener in mind was presented by Paraboni et al. (2007). They show how minimal descriptions, while logically distinguishing, can be unhelpful in a large spatial setting such as a university campus. Even if only one *room 198* exists on the whole university campus, a listener would arguably at least need the name of the building in addition to the room number, in order to be able to find it. Paraboni et al. present two algorithms that generate over-specified descriptions that are aimed at facilitating the search for the target referent in hierarchically structured domains such as university campuses or collections of structured documents. They showed in a task-based evaluation that in complex, hierarchically structured domains the over-specified descriptions produced by their systems are both preferred over logically minimal descriptions by human judges and help them to find the intended referents more easily. It might be noted that the logically minimal descriptions that Paraboni et al. aim to avoid seem unlikely to appear in a corpus of referring expressions produced by human speakers. They would therefore also be avoided by an algorithm that aims to replicate speakers' behaviour and not only by one specifically tailored to the needs of listeners. It would be interesting to see how the output of their algorithms would compare to human-produced referring expressions aimed at the same task in the same domain.

As discussed in the previous section, speakers do not always produce minimal descriptions, and Paraboni et el.'s work shows that this might be in the interest of listeners. However, other research in psycholinguistics has demonstrated that speakers are not always primarily concerned with making their listeners' lives as easy as possible when they build referring expressions (Horton and Keysar, 1996; Haywood et al., 2003, 2005). Investigating how it might be possible to replicate the reference behaviour of speakers, which is part of the aim of this thesis, is therefore not the same undertaking as building REG systems that produce referring expressions that are ideal for listeners, which should be the aim of systems to be employed in actual NLG applications.

2.5.3 Aiming for Naturalness

A third perspective on optimality in referring expression generation that is becoming more and more popular is the view that REG algorithms should generate descriptions that are as natural as possible and thereby account for human-produced referring expressions we find in corpora. The aim of generating natural-sounding referring expressions can be interpreted at three different levels which are somewhat akin to Chomsky's (1965) hierarchy of the adequacy of grammars:

1. *Operational adequacy*: The minimum that can be expected from an algorithm with the goal of generating natural referring expressions is that it never generates a referring expression that no human speaker would use in the same reference context.

2. *Descriptive Adequacy*: The second level of interpretation would require a REG algorithm to be able to generate all referring expressions that can be observed in human communication in the same context, and only those.

3. *Explanatory Adequacy*: The most ambitious goal in terms of naturalness involves more than just replicating human behaviour; it requires an algorithm to generate referring expressions *the way people do it*. An algorithm achieving this goal would provide an explanation of why humans produce the descriptions they do under a given circumstance.

Most algorithms have the implicit goal of producing referring expressions that sound 'natural', in the sense that a human would also use these referring expressions, and therefore subscribe to the first level of naturalness. Failing to do so might result in referring expressions that sound outlandish and would most likely either be very hard to understand or raise false implicatures in the listener about the intended function of the reference.

The aim for naturalness was first made explicit by Dale and Reiter's (1995) observation that many human-produced referring expressions do in fact contain information that is strictly speaking unnecessary to distinguish the target referent from the distractors, which means that minimal referring expressions are not always the most natural option. This led them to propose the IA, which abandons the goal of minimality in exchange for better computational tractability and more human-like output. Following this, researchers began to incorporate into their algorithms more and more phenomena that have been observed in human-produced referring expressions but were not replicable using the existing approaches. I discussed many of these approaches in Sections 2.2.2 to 2.3.; they include phenomena such as discourse salience (Krahmer and Theune, 1998, 2002; Theune, 2000), relations between objects (Dale and Haddock, 1991b; Krahmer et al., 2003; van Deemter and Krahmer, 2007), descriptions of plural entities (Gardent, 2002; Horacek, 2004; Gatt, 2007; Gatt and van Deemter, 2005; Areces et al., 2008), disjunctions and negations of properties (Gardent, 2002; van Deemter, 2002; Horacek, 2004; Varges, 2004; van Deemter and Krahmer, 2007), and adequate treatment of vague properties (van Deemter, 2000; Horacek, 2005; van Deemter, 2006).

Of course, just as the salience of properties and objects is likely to impact the ease with which a listener can resolve a reference, a speaker is also likely to prefer properties and landmarks that are particularly salient simply because they present themselves more readily for inclusion in a referring expression. Under the incremental model of language production proposed by Levelt (1989), a speaker might already have uttered the most salient properties before she finishes scanning the whole scene and determines which properties rule out any distractors. This means that the approaches mentioned in the previous section, which take visual salience into account and thereby alleviate the cognitive burden on the listener, are also likely to be better at mimicking the behaviour of speakers, whether this was an explicit goal of the algorithms or not.

Taking the step from the first to the second level of naturalness still does not provide a sufficient answer to the question of what should be considered an optimal referring expression, because people do different things in the same (referential) situation. Corpus studies, including those presented in the upcoming chapters, have shown that keeping all factors as equal as possible does not necessarily result in the same referring expressions being produced in every case. This inter- and intra-personal variation begs the question of what exactly we are aiming to model when our goal is human-likeness: all speakers, one particular speaker, the majority of speakers?

At least the aim of modelling *all* speakers, but most likely all of these aims, would expect an algorithm to produce a whole set of referring expressions for each target–domain pair rather than just one. This set would have to contain all referring expressions that any speaker might ever produce and only those. The problem with such an approach is that it is hard to know when we have found the complete set. Even if we have a collection of all descriptions ever given for a certain object in a certain situation, there is no guarantee that this collection includes the next referring expression that a speaker might produce.

The best we can do is to base our definition of optimality on the behaviour exhibited by *most of the speakers* that we are able to observe.[8] This definition of optimality is in line with the approach most psycholinguistic studies take: rather than attempting to cover all possible behaviours, psycholinguistic models aim at explaining and predicting a proportion of behavioural outcomes that is larger than what is likely to occur by chance. Of course our aim should be to build models and algorithms which cover as many speakers as possible, rather than being satisfied with a statistically significant majority. While fully attaining something analogous to Chomsky's explanatory adequacy seems like a rather ambitious goal, achieving descriptive adequacy for large corpora of human-produced referring expressions might get us one step closer to a model that in fact explains how people refer. This is the perspective that I will be taking on optimality in this thesis, in that the aim of the research I present is to further our understanding of how people refer and how computers may be able to replicate people's referring behaviour, rather than concentrating on the needs of the listener.

2.6 Discussion

I began this chapter with a description of what I take to be, for the purpose of this thesis, the task of a referring expression generation system: to select, at a purely semantic level, the properties which should be used to single out the intended referent from the distractor objects in the domain.

Following this, I discussed existing approaches to referring expression generation with a special focus on work that concentrates on content selection and on distinguishing descriptions. I described the different computational approaches that have been proposed for the generation of referring expressions. The guiding principles underlying the development of these approaches have shifted over

[8] Of course, this approach would equally make it possible to model the behaviour that one particular speaker has exhibited *most of the time*. This however, would require the collection of a large number of samples from the same speaker, which is more difficult to obtain.

the last 20 years. Early algorithms were influenced by an idealism striving for referring expressions of minimal length (full brevity search and constraint-based mechanisms); this ideal was then overtaken by a pragmatism which allowed some redundancy in the semantic content in order to reduce computational complexity (greedy and incremental search); and finally, much of the field has moved towards attempting to incorporate more and more of the phenomena that can be found in human-produced referring expressions, such as relations between entities, plural referents and disjunctions of properties (graph-based approaches and extensions of incremental search).

In the second part of this literature review I took a closer look at approaches that tackle a particularly difficult problem: the generation of referring expressions that make use of relations between the target referent and other objects. We will see in the following chapters that the majority of these approaches are not capable of generating the relational referring expressions that are present in the corpora I describe and use in this thesis. Most of these approaches involve fairly simple extensions based on recursive calls of existing algorithms that only work if relations are massively dispreferred compared to non-relational properties. The preference orders and cost functions that have been suggested to guide approaches to relational reference are usually not based on a sound empirical backing. Some make reference to psycholinguistic findings claiming that relations should be dispreferred due to their higher cognitive load. These psycholinguistic findings are rarely examined in detail and it is not clear whether they apply in the domains in which the algorithms are discussed, let alone in general.

In Section 2.4, I examined existing work in the field of REG that involves the use of human-produced corpora. I presented a number of existing corpora containing referring expressions and discussed the types of issues which can be studied in each of them. Many corpora are collections of dialogues with annotated referring expressions. In these corpora it is difficult to isolate the identification function of referring expressions from influences from the discourse context and conflicting functions arising from the larger task of the dialogue. The TUNA Corpus is closest in spirit to the type of data collections this thesis is based on, in that it contains referring expressions that were collected in isolation and have the sole function of allowing an onlooker to identify the target referent in a visual scene. However, the majority of instances in this corpus make reference to plural entities, which introduce new difficulties beyond the scope of this thesis. Also, the corpus is split over two very different domains, and two pragmatic design factors (fault-criticalness and the usefulness of spatial information) were varied in the collection exercise; these are likely to have a large influence on the content of the referring

expressions produced. In addition, there was almost no use of relations between objects in this corpus. To address some of these limitations I collected the corpora which I present in Chapter 5.

Following the discussion of existing corpora, I presented two existing corpus-based approaches to REG which automatically learned rules or decision trees from the data. Automatically learning algorithms from human-produced corpora has two advantages: it ensures that the resulting system is able to replicate the human data as well as possible based on the features chosen for learning, and at the same time allows the analysis of these features for their usefulness based on the accuracy scores that are achieved with or without each feature. In addition, decision trees and rule sets have the particular advantage of being easy to inspect, which permits post hoc analysis. One of the existing systems was aimed at explaining the content of referring expressions in a corpus of human dialogues based on the varying functions the referring expressions fulfilled, and the other one predicted the form rather than the content of referring expressions. In Chapter 6 I will present a machine learning approach to referring expression generation that is solely aimed at content selection for distinguishing descriptions.

Finally, I described two evaluation experiments in which the output of existing algorithms was compared to corpora of referring expressions. I pointed out a number of shortfalls in both of these studies. In particular, they both evaluated only a small number of possible parameter settings for each algorithm. This procedure makes it impossible to draw conclusions about the capabilities of these algorithms to generate the referring expressions found in the corpora *in principle*. So, these studies only shed limited light on the question of whether or not these algorithms can replicate human data. I will address this issue in my evaluation experiment in Chapter 4.

The last section of this chapter examined three popular interpretations of what it means for a referring expression to be optimal. The first of these was maximal brevity based on the Gricean Maxims, which has had a small revival in recent years, despite psycholinguistic evidence showing that it is not a good model of what speakers do and also often not what listeners need. The second interpretation takes a listener-oriented perspective, which has been incorporated in a number of approaches that attempt to make it easier for the listener to resolve a referring expression by including properties that are easy to process. The third interpretation of optimality takes the opposite, speaker-oriented, perspective: instead of attempting to make the listener's job easier, which makes sense for systems that are designed for specific applications, approaches taking this perspective try to model what speakers do and thereby further our understanding of the cognitive processes involved. This is the perspective that I take in this thesis.

Chapter 3

Methodological Choices

In this thesis, corpora of referring expressions are involved in three tasks: corpus collection, semantic analysis, and gold standard evaluation. In this chapter, I discuss the methodological choices and conceptual decisions that have to be taken before any of these three tasks can be tackled.

3.1 Issues in the Collection of Corpora

When collecting a corpus of referring expressions, or any natural language corpus for that matter, a number of decisions as to the form this corpus will take are influenced by the aim of the exercise. If this aim is ultimately to develop algorithms that can mimic and maybe even explain to some extent what humans do, based on what we observe in a corpus, we have to ensure that the referring expressions in this corpus are as natural as possible, in the sense that they actually might be used by a human in a natural reference context. As with almost all psychological and psycholinguistic experiments, it turns out that achieving this goal requires a balancing act between ensuring maximal control over the factors that might be influencing the outcome of the experiment (in this case the content of referring expressions) and maximal naturalness of the situation in which the data is collected.

3.1.1 Collected vs. Found Data

Arguably the best way to ensure that the referring expressions in a corpus are going to be natural is to collect them in real-life situations. For example, one could annotate the referring expressions found in newspaper text or transcribed naturally occurring speech. These referring expressions would be guaranteed to be of the kind we want to mimic and explain.

However, the real contexts in which referring expressions are used can be very complex. Consider the following hypothetical references:

(3.1) Turn left after *the second shopfront that has a 'For Lease' sign in the window.*

(3.2) Do you mean *the keys that are under the loose leaf folder on the desk in the upstairs study*?

(3.3) *The member for Keira* apologised to his family, the Premier and his constituents for his actions.

In real life situations giving rise to referring expressions such as these, there are generally too many variables to permit carefully controlled experiments that would allow us to derive general principles for content determination.

Using a corpus of 'found' referring expressions from 'real' text affords almost no control over the circumstances under which the referring expressions were produced beyond, perhaps, the choice of text type and genre. A myriad of discourse factors might have had an impact on the generation of such purely natural referring expressions, such as global and local goals of the discourse, register, expected knowledge level of the reader or listener, different referring functions, and so on. Determining these factors would be extremely hard. Even pinning down the necessary reference context for the generation of a referring expression with a purely distinguishing function would be difficult in such a corpus. It would be almost impossible to determine exactly which were the distractors that the target referent was distinguished from, what were the properties of these distractors that were taken into account, and what was the original set of properties of the target referent from which the choice for the referring expression was made.

It seems, therefore, that in order to gain some level of control over these factors, we need to relinquish the highest level of achievable naturalness by collecting corpora in systematically designed settings. These settings need to be designed to allow us to restrict the aspects of the reference situation that might be impacting on the content of a referring expression in a principled way, while being as close to natural reference situations as possibile. In line with almost all work in this area (see, for example, Brennan and Clark, 1996; Thompson et al., 1993; Jordan and Walker, 2005; Gatt et al., 2007), I therefore carry out my explorations in rather simple scenarios that allow me to explore specific hypotheses and to characterise the general strategies that humans seem to adopt; these strategies might then be deployed in more complex scenarios to see whether they continue to be applicable.

A second disadvantage of corpora of referring expression found in natural text is that we are unlikely to find many references to the same entity by different

speakers. This makes it impossible to study cross-speaker variation, which, as we will see, is a major factor in the generation of referring expressions.

3.1.2 Reference in Discourse vs. Isolated Reference

The question of whether to collect referring expressions in a discourse context or whether to elicit isolated referring expressions without such a context is related to the choice between collected and found data discussed in the previous section. Of course, most reference occurs as part of a larger discourse rather than just out of the blue as a one-off event, which means that referring expressions obtained in discourse will be more likely to fulfil our naturalness criterion.

However, discourse, and especially dialogue, introduces a large number of factors which are hard to control and at the same time likely to have an impact on the content of a referring expression. This is the case in naturally found discourse, as argued above, but also in discourse whose aim and content has been controlled to some extent by experimental settings. Examples of this are the COCONUT dialogues and the Maptask and iMap dialogues, mentioned in Section 2.4.1. All of these corpora were collected in controlled settings; but nonetheless there are a large number of factors influencing the form and content of the referring expressions contained in them that go beyond the requirements arising from the function of identification. In the COCONUT Corpus in particular, the properties contained in referring expressions often carry out additional functions such as convincing the dialogue partner of a certain course of action or dissuading them from buying a specific furniture item.

Even in corpora where identification of landmarks is the main function of reference, discourse context can complicate the matter considerably. A large body of psycholinguistic evidence shows that, over the course of a dialogue, the form and content of referring expressions are shaped to a large degree by processes such as **naming**, **alignment**, **co-ordination** or **co-operation** (see, for example, Krauss and Weinheimer, 1964; Carroll, 1980; Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Wilkes-Gibbs and Clark, 1992; Brennan and Clark, 1996; Branigan et al., 2000; Haywood et al., 2003; Metzing and Brennan, 2003; Pickering and Garrod, 2004). What all of these processes imply is that referring expressions change as a result of speakers (partly) mimicking the way things have been described before and taking into account needs of the listener that might have come to light during the conversation.

It is interesting, worthwhile, and even important, to explore these processes, and eventually mimic them in NLG systems if these systems are to be employed

in discourse settings or if they are intended to model human discourse processes. However, a systematic approach to understanding distinguishing reference, has to necessarily begin by attempting to model reference in isolation, before we might hope to be able to fully model all of these factors in parallel. The corpora I explore in this thesis therefore contain isolated referring expressions, which were collected in a consistent, strictly controlled reference context.

This is not to say that referring expressions such as the ones in my corpora, or the TUNA Corpus, never occur in more natural discourse. The reference situation in which they were collected is similar, for example, to that in which the first descriptions of landmarks in direction-giving occur. These referring expressions usually fulfil the sole purpose of identifying the target referent, and as long as we do not concern ourselves with subsequent mentions of the same referent, we can assume that any influence of additional factors on the content of referring expressions is minimal.

3.1.3 Characteristics of Domains

One of the most important and most difficult decisions to make before embarking on a data collection exercise for a corpus of referring expressions is the choice of domain. I will discuss in this section three important decisions that have to be made as part of this choice.

The first decision to be taken is concerned with the presentation of the entities to be described. While REG systems can be used in non-visual domains, such as for example Dale's (1989; 1992) system for generating recipes or Siddharthan and Copestake's (2004) algorithm for re-generating references from text, it is difficult to present stimuli for the generation of distinguishing descriptions in a non-visual way without already describing the objects and thereby priming the participants to reuse the same descriptions. One option is to allow people to describe things from memory, for example by prompting them to describe a famous monument or building or to give directions from one place to another by using landmarks.

The advantage of presenting subjects with visual stimuli is that it affords tighter control over the underlying knowledge that is needed for describing the target referent. In effect, visual stimuli, or a propositional representation of them, can be treated as the knowledge representation model of the speaker and the listener. This technique was used in the MapTask (Anderson et al., 1991) and iMap (Guhe and Bard, 2008a) experiments where two participants were presented with slightly different maps of the same environment. Two people having different maps of the same area might seem like an unlikely scenario, but it becomes more likely

if the maps are regarded as the slightly different memory that the interlocutors have of the same area. Instead of having to rely on different memories that people might really have of a place — which would be very hard to discern — the maps permit exact knowledge about the differences between the two speakers' knowledge representations.

In the data gathering experiments I describe in Chapter 5, stimuli are presented visually and the assumption is that both speaker and listener can see the same scene from roughly the same point of view. Assuming that both interlocutors can perceive the same information about the stimuli at the time that the referring expression is uttered ensures that the referring expressions function purely as identification devices rather than as gaze-directing devices. Directing referring expressions can occur when the speaker assumes that the listener is currently not looking at the target referent and needs to be directed towards it. Description (3.4) is an example of such a directing referring expression.

(3.4) the button on the wall behind you

These types of referring expressions are common in direction-giving systems such as the ones being evaluated in the GIVE Challenge, where an NLG system has to help a human player navigate a virtual environment (Byron et al., 2009; Koller et al., 2010).

A second choice about the domain concerns the nature of the objects to be described. This choice again turns out to require a trade-off between naturalness and experimental control. In the TUNA Corpus, drawings of furniture items and photographs of male mathematicians were used. Clearly, photographs are a more natural type of stimulus than stylised drawings of furniture items that have been retrospectively coloured in one primary colour. However, Gatt and colleagues (Gatt et al., 2007; van der Sluis et al., 2007; Gatt, 2007) note that the part of the corpus based on photographs is much more complex in terms of the variation in people's descriptions and was much harder to annotate consistently. This variation was due to a much larger number of features in the photographs than in the drawings and to participants even finding 'new' features to describe the mathematician's faces that had not been recorded in the knowledge base. To ensure maximum possible control over the properties that people are likely to use in their referring expressions, I opted for scenes with rather simplistic objects (spheres and cubes) in my corpora, rather than real or naturalistic objects.

This was offset by the third decision, which was to use a 3D presentation affording a more natural spatial representation than the 2D grid displays used for the TUNA Corpus and the drawer corpus which I used for the evaluation in

Chapter 4. The choice between 2D grids and 3D scenes has most bearing on the use of spatial properties. In 3D scenes the spatial relations between the objects are more natural, while in a 2D grid the location of an object can be pinned down more exactly in terms of co-ordinates, which makes it possible to distinguish objects not only by their inherent properties but also by their position in the grid. In many cases, grids are an artificially imposed representation of space which is not appropriate in everyday situations. There are, of course, some natural domains with grid-like properties, such as car park layouts or the filing cabinet domain we will see in Chapter 4, which make the study of reference in 2D grids a worthwhile undertaking in itself.

3.1.4 Web-based vs. Off-line Data Collection

The internet makes it possible to conduct data collection experiments online, rather than having to recruit participants locally and physically interacting with them. Web-based data collection has a number of significant advantages over off-line experiments:

- It decreases the time-constraints on the experimenter, as there is no need to supervise the participants while they do the experiment. On the other hand, web-based procedures are, of course, not suitable for experiments that require personal supervision by the experimenter.
- It makes participation in the experiment more convenient for the participants, who can choose any time and place convenient to them, rather than having to make an appointment with the experimenter.
- The two points above result in the possibility to collect a larger quantity of data from each participant and to collect data from a much larger number of participants.
- Web-based experiments make it much easier to recruit participants from a large variety of backgrounds and age groups than what is usually the case in experiments carried out in a university laboratory.
- The more accustomed people become to communicating and interacting online, the more natural the web becomes as an experimental environment compared to a lab setting, where participants might feel supervised and more under pressure to give 'correct' answers.

At the same time, using the web as a platform for conducting experiments means that some experimental control has to be relinquished. While it is easy

to collect demographic information from the participants in a web form before or after the experiment, there is no way to confirm the accuracy of this information. Not being in the presence of the examiner might lower the threshold for participants to lie about their age or give bogus information about themselves just as a joke. At the same time, the anonymity of the web-based setting might make people less reluctant to disclose their real information. In the context of the data collection experiments described in this thesis, the only piece of personal information of real importance is whether the participants were native speakers of English or not. However, in experiments that require data from different age groups, social demographics or professional backgrounds, the lack of real control over these factors might make a web-based setting not ideal. A second factor which cannot be controlled effectively over the web is how seriously the participants take the exercise. They might not read the instructions carefully, they might lose interest during the experiment and abandon it for a computer game, or they might start giving unrelated responses. In a language production experiment, it is fairly simple to recognise and filter out unrelated responses and compensate by collecting more data. It might even be possible to recognise data that was produced as a result of not fully understood instructions. The same might not be the case in other types of experiments; but some preliminary evidence shows that annotation data collected online is no less reliable than data from off-line experiments (Snow et al., 2008).

3.2 Issues in the Analysis of Corpora

In this section, I discuss a number of concepts that play an important role in the analysis of the referring expressions contained in a corpus. Of course, the first step of a corpus analysis exercise aimed at semantic content must be to derive this semantic content for each expression in the corpus. After all, human participants do not generate sets of properties, as most REG algorithms do, but rather fully realised noun phrases. The methods that I used to derive the semantic content from the human-produced noun phrases will be detailed in the appropriate sections where the corpora are described. Once this step has been performed, the analysis can concentrate on the attributes contained in the referring expressions (see Section 3.2.1) and characteristics of the referring expressions that arise from the presence or absence of individual attributes (Sections 2.5 and 3.2.2).

3.2.1 Types of Object Attributes

We already saw in Section 2.2.2 that not all attributes are the same, as has been noted, in particular, by van Deemter (2000, 2006). In the context of the corpora discussed in this thesis, we will encounter four different kinds of attributes: object types, absolute attributes, relative attributes, and spatial attributes including relations and location attributes. We will see in Chapters 5 and 6 that the differences between these kinds of attributes result in different usage patterns in human-produced referring expressions.

The type of an object constitutes a special case because it is very rarely omitted from a referring expression, and this is the case in all human-produced corpora under discussion in this thesis. Consequently, most algorithms treat type separately to ensure that it is added to every referring expression. One partial explanation for this special status is that referring expressions get realised as noun phrases, every noun phrase requires a head noun, and it is usually the referent's type that gets realised as the head noun. I will discuss the special status of type in more detail in the context of a definition of what it means for a referring expression to be minimal in Section 3.2.2.

van Deemter (2000, 2006) has argued that there is an important difference between attributes whose values are always absolute, such as colour, type, or even names, and those which can have vague or gradable values, such as size. Of course, size can have absolute values, expressed, for example, in millimetres; however, in the majority of cases that we encounter in corpora such as the ones used in this thesis, their values involve a degree of relativity to the size of other objects: an object is only *tall* in comparison to another shorter object. We will see that the usage pattern for size in my corpora provides support for Brown-Schmidt and Tanenhaus's (2006) hypothesis that, differently from other, absolute, properties such as colour, size only gets used if a distractor object is present that is of the same type.

The spatial attributes of the objects in the simple visual contexts used for the corpus collections in this thesis, as well as in most other contexts used in REG research, fall into two categories: (1) the location of an object within the scene, such as the one in Example (3.5), and (2) spatial relations between an object and another nearby object, such as the one in Example (3.6).

(3.5) the ball *in the left*

(3.6) the ball *on top of the blue cube*

Of these, location has occasionally been classed as a vague attribute, similar to size

(Gatt et al., 2007). After all, a property such as in_the_left is not an exact position in terms of pixels or even a grid reference.

Relations between the referent target and another entity, such as the one between the ball and the cube in Example (3.6), pose a big challenge for referring expression generation algorithms because they involve not just one value for an attribute, but rather a whole new object with all its properties and relations which all could be included in the referring expression. In Section 2.3, we saw a variety of approaches that tackle the task of generating referring expressions that contain relations between objects. In Chapter 4, I will discuss why none of these approaches would be able to generate the particular relational descriptions found in the corpus under discussion there, and in Chapter 5, we will see that the assumptions made by most existing approaches to the generation of relational referring expressions do not hold for the two corpora I analyse there.

3.2.2 Minimality and Over-specification

The concepts of minimality and over-specification play an important role in the analysis of human-produced referring expressions. As discussed in Section 2.5.1, one popular view taken on what it means for a referring expression to be optimal is that a referring expression should be as short as possible while fully distinguishing the target referent from all distractor items. The shortest possible referring expression for a target referent in a given domain is usually called the **minimal distinguishing description**. A referring expression that contains more information than necessary to distinguish the intended referent from the other objects in the context set is often called **over-specified**, and the unnecessary properties contained in it are called **redundant**[1].

In the following, I take a closer look at the definitions of minimality and redundancy. I first examine the definitions of minimality that are provided in the literature and will then move on to finding a definition for over-specification.

Minimality According to Dale (1989)

Dale (1989, p. 71), the first proponent of minimality in REG, defines a minimal distinguishing description (MD) as

> a set of such attribute–value pairs, where the cardinality of that set is such that there are no other sets of attribute–value pairs of lesser cardinality which are sufficient to distinguish the intended referent.

[1] Confusingly, the term *redundant* has sometimes been used for referring expressions as a whole, where the term *over-specified* should have been used instead. This is something of which I am guilty myself (Viethen and Dale, 2006a).

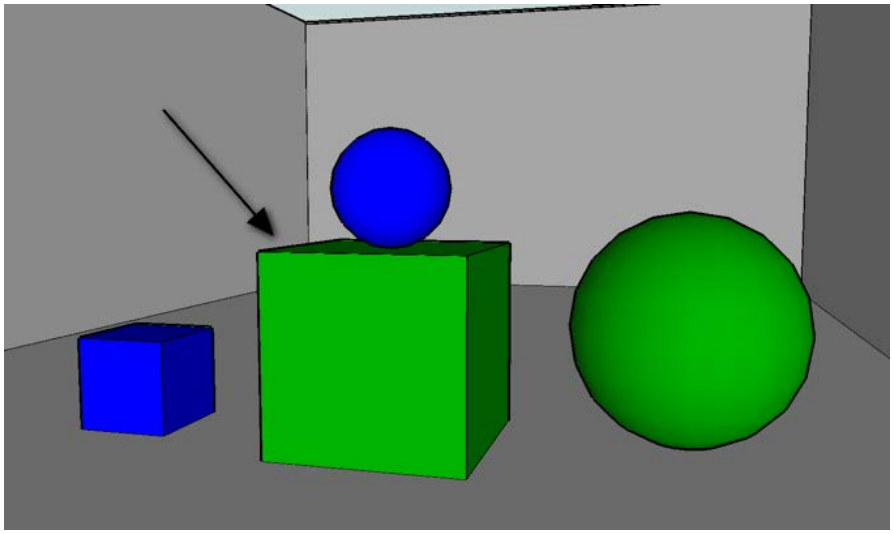

Figure 3.1: A large green cube

There are two things worth pointing out in this definition. Firstly, Dale defines the length of a description in terms of the number of attribute–value pairs it contains, that is, the amount of semantic content contained in a description, rather than its length in number of words, characters or actual time taken to utter. For analyses at a purely semantic level, such as mine, this definition is the only useful one as there are usually different possible realisations for the same semantic content. However, for work at, for example, lexical, syntactic or phonetic levels, it might be more appropriate to adopt definitions of length that take into account the number of characters or words.

Secondly, Dale writes that a minimal description is *a*, not *the*, set of attribute–value pairs, acknowledging the fact that there might well be more than one minimal description in any given situation. For example, the target referent in Figure 3.1 (marked by a black arrow) can be described as in Example (3.7) or as in (3.8). Both of these descriptions contain the minimum number of properties necessary to distinguish the target from the other three objects in this scene, and none of these could be left out without rendering the description ambiguous.

(3.7) the large cube {⟨size:large⟩, ⟨type:cube⟩}

(3.8) the green cube {⟨colour:green⟩, ⟨type:cube⟩}

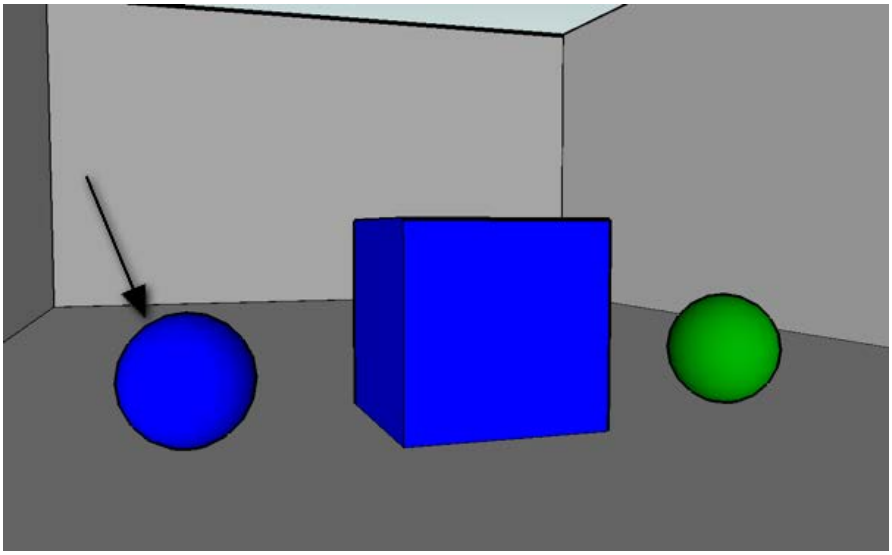

Figure 3.2: A small blue ball.

Minimality and type

There are, however, a number of problems that arise from Dale's definition. The first complication that we encounter in this definition is due to the fact that most REG algorithms always include a type property. They do this based on (1) the deliberation that the following realisation step will require a property that can be realised as the head noun of the description, and (2) the fact that in most human-produced referring expressions the head noun is a realisation of the type property. This decision is supported by findings in psycholinguistics which demonstrate that humans often include type even when it does not add any discriminatory power to the description (e.g., Pechmann, 1989). In the minimal descriptions (3.7) and (3.8), the type cube was required for distinction, so this did not cause any difficulty; but consider the scene shown in Figure 3.2 where the smallest sets of properties by which the intended referent can be distinguished are shown in Examples (3.9) and (3.10).

(3.9) {⟨size:small⟩, ⟨colour:blue⟩}

(3.10) {⟨colour:blue⟩, ⟨type:ball⟩}.

In Example (3.10) the type attribute is included as a distinguishing property and cannot be dropped. Example (3.9), on the other hand, does not contain type. Virtually every implementation of REG algorithms (e.g. Dale and Reiter, 1995; Kelleher and Kruijff, 2005, 2006; Viethen and Dale, 2006a; Gatt et al., 2007;

van der Sluis et al., 2007) would additionally also include ⟨type:ball⟩ in the set of properties shown in (3.9) in order to facilitate the realisation of a head noun. Without knowing exactly how an algorithm works, it is impossible to discern from the outcome whether the type property was added to the description simply to aid realisation or for some other reason, for example, because the circumstances warrant an over-specified description. Should we then consider the property set in (3.11) over-specified because size could be dropped? Or is it, in fact, a minimal description to which type was added retrospectively because a head noun was needed?

(3.11) {⟨size:small⟩, ⟨colour:blue⟩, ⟨type:ball⟩}

If we cannot determine the answer to this question for descriptions produced by algorithms which we know always include type, it is impossible to answer it for human-produced referring expressions without postulating one exactly specified algorithm as the one that people employ for the generation of referring expressions.

Based on Kempen and Hoenkamp's (1987) head-driven model for syntactic construction and what he calls Pechmann's Gestalts Principle (Pechmann, 1989), Gatt (2007) suggests a **category-driven approach to reference** (p. 125) which is centred around the type of the referent. The category-driven approach to reference always includes a type attribute based on the hypothesis that it constitutes the core of an object's conceptual gestalt. Consequently, type should always be the first property to get included in each referring expression, rather than being added at the end if it is still missing. In practice, it frees the content determination process from its responsibility to present to the realisation process a property that is realisable as a head noun, because it always includes one such property anyway. Under a strict interpretation of this model, a description without a type property such as the one in Example (3.9) could never occur, and (3.11) has to be counted as (intentionally) over-specified, because colour and size were both added in the full knowledge that type had already been chosen first and that, therefore, the shorter (3.10) could have been produced instead.

This analysis works well for the corpora used in this thesis, as all objects are simple enough to make it trivial to determine the basic level type (in the sense of Dale and Reiter's (1995) class hierarchy, c.f. Section 2.2.2) in each case and it is always the expected type property that gets realised as the head noun. However, this does not always have to be the case. In a domain where all or most objects are of the same type, people might drop the type property and instead pick a different property to be realised as the head noun. For example, in the alien domain of the iMap corpus (Guhe and Bard, 2008a), different aliens were distinguishable only

by their colour and their shape. This leads some speakers to produce descriptions such as Example (3.12) instead of (3.13).

(3.12) the green rectangle

(3.13) the green rectangular alien

Arguably, the speaker who uttered (3.12) conceptualised the target referent as a rectangle rather than an alien, which means that the category-driven approach would analyse it as containing the attribute–value pairs in (3.14), rather than those in (3.15) with a type missing. This leads to a new question regarding the definition of minimality: Is a description such as (3.13) minimal even though it would have been possible to find a shorter description such as (3.12) under a different conceptualisation? Although interesting in principle, I will not discuss this issue further as it has no bearing on the analysis of my corpora.

(3.14) {⟨colour:green⟩, ⟨type:rectangle⟩}

(3.15) {⟨colour:green⟩, ⟨shape:rectangle⟩}

Minimality and Locational Properties

We have already seen that several equally short minimal descriptions can exist for the same target referent in the same context. So far, the examples I have used did not contain any information about the location of the target referent, although the visual domains in the example scenes make the use of such spatial information entirely possible. For example, the target referent in Figure 3.2 could also be described by the referring expression in Example (3.16).

(3.16) the blue ball in the left

Such spatial information increases the number of different distinguishing descriptions that can be used for an object, and it is possible that locational descriptions exist which are shorter than, or just as short as, the shortest description that does not use location.

Dale (1989) did not have to deal with such cases as his domain was non-visual, making locational information useless. Gatt, on the other hand, discussed the difference between locational information and what he calls **inherent visual attributes** in the contexts of minimality and redundancy (Gatt, 2007, pages 65–67 and 78–80). I will adopt this term in order to refer to properties such as colour, type and size as opposed to spatial properties such as location or spatial relations to other objects. His definition of a minimal description does not allow any spatial

information to be included. He postulates that only properties that form part of the immediate gestalt of an object can be part of a minimal description, and called these properties 'MD attributes'. This does not include location, which is an external attribute whose use, in Gatt's view, stems from a different conceptualisation of the target referent.

The view that locational attributes are qualitatively different from inherent visual attributes is surely correct. The inherent attributes are those that stay unchanged, no matter in which context an object is viewed. Locational attributes on the other hand change, once the object is moved to a new environment. However, I do not agree that this precludes locational or, in fact, relational information from the group of attributes that can appear in minimal descriptions. Arts' (2004) studies show that locational information, including relations to other objects, is highly preferred by listeners and often used by speakers. This makes sense intuitively, as the spatial location of an object within a visual scene can arguably be a highly salient property: corner positions, central or extreme peripheral positions as well as spatial relations that occur only once in a scene are all properties that are likely both to be very useful in a referring expression and to require little cognitive effort to recognise.

What Gatt calls a minimal description (MD) is not actually the shortest possible referring expression but, in fact, the shortest possible referring expression using only inherent visual attributes. I will therefore call this type of description a **inherent minimal description** (inherent MD). Given the qualitative difference between inherent and locational attributes, it is nonetheless worthwhile to track the occurrence of inherent MDs in a corpus.

Four Definitions around Minimality in Referring Expressions

Based on the above discussion, I adopt the more general definition of a minimal description given by Dale (1989). Following the category-driven approach to reference, I assume that every referring expression contains a property that represents the type of the target referent and that type is always included as the first attribute. My definition of a minimal description is then as follows.

Definition: Minimal Description (MD) — A set of properties true of the target referent such that no other fully distinguishing set can be found that is smaller. type is always counted towards the set of properties contained in a referring expression.

Locational properties and relations can be part of an MD; however, I define the following three subtypes of an MD:

Definition: Locational MD — An MD which contains at least one locational property. Each locational property is counted separately (e.g. if in the left and in the top are contained in a referring expression, they count as two properties.).

Definition: Relational MD — An MD which contains a relation between the target referent and a landmark object. A relation counts as one property towards the length of a referring expression. Each property of a landmark object included via a relation is counted separately.

Definition: Inherent MD — An MD that contains neither a locational attribute nor a relation to another entity.

Over-Specification in Referring Expressions

An over-specified referring expression is one that contains at least one property that is redundant. Based on the same reasoning that resulted in counting type towards the length of a referring expression, namely that type is included not primarily to distinguish but rather because it is the core of the referent's gestalt, I will consider a referring expression to be over-specified even if only its type is redundant. Therefore, my definition of an over-specified description is as follows.

Definition: Over-specified Description — A description that contains at least one redundant property. type being used redundantly does also make a referring expression over-specified.

The Euler diagram in Figure 3.3 visualises the relationships between the different types of referring expressions for a given target referent.[2] Each referring expression is either distinguishing (green sets) or it is under-specified (orange set). The set of distinguishing descriptions has two subsets: a distinguishing description is either over-specified (dark-green set), or it is not. Gatt (2007) calls the set of distinguishing descriptions that are neither under- nor over-specified 'well-specified'. I will not use this term because it seems to imply that it is better for a referring expression not to be over-specified. Where I need to refer to a referring expression of this type I will call it a **sufficient distinguishing description** (light-green sets in the diagram). A member of the set of sufficient distinguishing descriptions

[2]Of course, the proportions in this representation are not necessarily indicative of the sizes of the sets. In fact, the sizes of the sets will differ for each reference scenario.

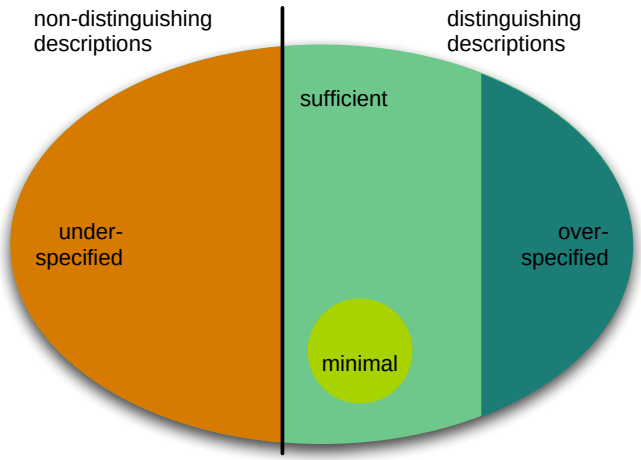

Figure 3.3: A visualisation of the relationships between over- and under-specification

can additionally be a member of the subset of minimal descriptions (circle-shaped yellow-green set).

To illustrate these different types of referring expressions, let us consider an example scenario where the target referent is a table in a furniture showroom that contains a whole range of different tables. If it is the only green table in the room, Description (3.17) is a minimal description for it. Description (3.18) might be another distinguishing description for the same table. It is clearly not minimal as it is longer than (3.17), but it is not over-specified if there is at least one other long hexagonal table, one other hexagonal table with a bowl on it and one other long table with a bowl on it, but no other table possessing all three of these properties. (3.18) is therefore a sufficient, but not minimal, distinguishing description. It is important to note that it is possible for an under-specified description to be longer than a distinguishing description. For example, Description (3.19) is under-specified because there is another long hexagonal table. Yet, this description is longer than the distinguishing description (3.17). Similarly, over-specified descriptions are not necessarily longer than all other distinguishing descriptions. Description (3.20) is over-specified because the property long is redundant, but this description is shorter than (3.18).

(3.17) the green table

(3.18) the long hexagonal table with the bowl on it

(3.19) the long hexagonal table

(3.20) the long green table

3.3 Comparing System Output to Corpus Data

Evaluating the output of a content selection algorithm for REG against corpus data involves comparing each referring expression generated by the algorithm to one or more corresponding referring expressions in the corpus which were generated for the same target referent in the same context. In this type of evaluation scenario, the referring expressions from the human-produced corpus are considered to be the **gold standard** that the system is aspiring to. The comparison between system output and gold standard can take place at a number of different levels of linguistic granularity, and a number of different comparison metrics have been used in recent REG evaluation exercises. In the following, I discuss these different options and indicate which ones I use in the later chapters of this thesis.

3.3.1 Common Evaluation Metrics

The simplest way to compare the content of two referring expressions to each other is to determine whether they are a perfect match or not. The percentage of referring expressions in a corpus for which the system can produce a perfect match represents the **Accuracy** achieved by the system. This has also been called **Recall** (Viethen and Dale, 2006a) or **Perfect Recall Percentage** (PRP) (Belz and Gatt, 2007; van der Sluis et al., 2007) in the literature.

Applied at the level of whole referring expressions, Accuracy is a relatively coarse-grained and consequently strict comparison metric. It assigns the same penalty to a system for producing a referring expression that omits one property but is otherwise identical to the gold standard as for producing a referring expression that has no semantic overlap with the gold standard at all. In some situations, it might be preferable to give a system partial credit for getting some or most of the properties right. One way of doing this is to use Accuracy at the level of attributes rather than complete referring expressions. In a sense, a REG algorithm can be viewed as a prediction system that predicts for each attribute of the target referent whether it should be included in a referring expression or not. The **Attribute-level Accuracy** for a whole referring expression can be reported as the number of the correctly predicted attributes in that referring expression or, if the length of the referring expression is to be taken into account, as the proportion of correctly predicted attributes. The overall Attribute-level Accuracy over a whole data set can then either be reported as the overall number of attributes predicted correctly or as the average of Attribute-level Accuracies the system achieved for the referring expressions contained in the data set. Machine learning systems use Accuracy as the default evaluation metric, and it is therefore Accuracy and Attribute-level

Accuracy that I report in the machine learning experiments in Chapter 6.

Another way of comparing system output to gold standards at a more fine-grained level than Accuracy is to use existing set comparison metrics such as the DICE coefficient (Dice, 1945; Salton and McGill, 1983) or MASI (Measuring Agreement on Set-valued Items: Passonneau, 2006). DICE was first used in REG by Gatt and colleagues (Gatt, 2007; Gatt et al., 2007; van der Sluis et al., 2007) and has since become something of a standard evaluation metric for content selection in REG. MASI was first proposed as a metric to measure agreement between annotations of Summary Content Units in the context of evaluation of automatic summarisation systems. It has been used in parallel with DICE in a number of evaluations including the later REG evaluation exercises (Belz and Gatt, 2008; Gatt et al., 2008; Viethen et al., 2010).

Both the DICE coefficient and the MASI score are set-comparison metrics that deliver values ranging between 0 and 1. In the context of content determination in REG, they are applied by comparing the set of properties contained in the description that the system has produced to those contained in the human-produced description. The main difference between them is that MASI is biased in favour of solutions that are a subset or a superset of the gold standard.

Given two referring expressions, represented by the sets of properties contained in them, A and B, DICE is computed as

$$\text{DICE(A,B)} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{3.21}$$

and MASI as

$$\text{MASI(A,B)} = \delta \times \frac{|A \cap B|}{|A \cup B|} \tag{3.22}$$

where δ is a monotonicity coefficient which implements the bias for subsets and supersets of the gold standard. It is defined as

$$\delta = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ 1 & \text{if } A = B \\ \frac{2}{3} & \text{if } A \subset B \textit{ or } B \subset A \\ \frac{1}{3} & \text{otherwise} \end{cases} \tag{3.23}$$

3.3.2 Attribute-Level vs. Property-Level Evaluation

In REG, Accuracy, DICE and MASI are often applied at the level of attributes rather than their values. This is done to avoid penalising systems for not replicating unorthodox attribute values differing from the values found in the knowledge base: for example, if a human participant who contributed to a gold standard called an object *egg-shaped*, but in the knowledge base that object has the property ⟨shape:oval⟩. This practice ensures that the evaluation takes into account only the semantic content of a referring expression and not the lexical choices that were made. It can be problematic for two reasons: First, some REG systems incorporate mechanisms to choose the best value for some or all properties, for example based on the expertise of the listener (Janarthanam and Lemon, 2009) or based on the assumption that using a more specific attribute value than necessary carries unintended implicatures (Dale and Reiter, 1995). If only the presence or absence of attributes are taken into account rather than the actual attribute values, the choices these mechanisms make cannot be evaluated.

The second reason that makes attribute-level evaluation problematic compared to property-level evaluation is that the choice to include an attribute in a referring expression is in most algorithms based on its value, because the discriminatory power of an attribute depends on how common the value for this attribute is. For example, in a domain with a number of blue objects and some objects in slightly different shades of red, a speaker might include ⟨colour:red⟩ to distinguish one of the red objects from all the blue ones and then choose another property to distinguish from the remaining red distractors. A REG algorithm might instead choose ⟨colour:dark-red⟩, which distinguishes from all objects at once and therefore has high discriminatory power, although this fact might only be perceivable to a human onlooker after thoroughly inspecting the colours of all red objects. Should this algorithm then be penalised for not choosing the same colour value as the human speaker, or rewarded for correctly including colour? To avoid having to deal with cases such as this, existing test domains and corpora maximise the differences between possible values of the same attribute. In other words, we will not find objects in the same scene that can be distinguished from each other by calling one *red* and the other *dark red*.

van Deemter and Gatt (2007) discuss these shortcomings in the context of the DICE co-efficient and come to the conclusion that it might be necessary to validate automatic, speaker-oriented evaluation, which compares system outputs to human-produced corpus data, by combining it with listener or user-oriented

evaluation, which involves humans rating the system output or performing a task, such as picking out the correct target referent based on the system output. This solution is of course only applicable if it is indeed our goal to build an application that is maximally useful for listeners. If, as is the case in this thesis, our goal is to replicate and explain human behaviour, speaker-oriented evaluation is the only option.

3.3.3 Taking Length into Account

van Deemter and Gatt mention another flaw of the DICE metric, which it shares with MASI: both metrics punish a description that is too short more harshly than one that is too long. Suppose the gold standard description contains the attribute set $G = \{P, Q\}$. A system-generated description omitting one of these properties, for example $S_1 = \{P\}$, would achieve $\text{DICE}(G, S_1) = \frac{2}{3} = 0.\overline{6}$ and $\text{MASI}(G, S_1) = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9} = 0.\overline{4}$. A description adding an extra property, for example $S_2 = \{P, Q, R\}$, would achieve $\text{DICE}(G, S_2) = \frac{4}{5} = 0.8$ and $\text{MASI}(G, S_2) = \frac{2}{3} \times \frac{4}{5} = \frac{8}{15} = 0.5\overline{3}$. van Deemter and Gatt suggest using a metric that, similarly to edit distance, punishes deletions and additions in the same way.

Usually, the quality of a REG algorithm is not assessed based on just one referring expression but on a whole data set containing many different referring expressions from slightly different reference contexts. The common way to extend DICE (and MASI) from a single referring expression to a whole set is to simply report the mean score that a system achieved over all referring expressions in a set. The simple mean, however, treats the scores for all referring expressions in a corpus the same, irrespective of the differing lengths of the expressions. It might be desirable to give the scores for longer referring expressions more weight in this overall score, to acknowledge that it is harder to replicate all properties of a long referring expression correctly than to do the same for a short referring expression. In (Viethen et al., 2010) we use what we called the **summed** DICE score, which gives longer referring expressions more weight in the same way as the BLEU metric used for Machine Translation gives more weight to longer sentences. It sums the denominators and the numerators of the individual DICE scores separately and then divides the numerator sum by the denominator sum, rather than taking the simple mean of all scores.

For example, let's assume system A achieved the DICE scores of $\frac{6}{6}$=1.0, $\frac{3}{10}$=0.3 and $\frac{1}{5}$=0.2 on a (very small) corpus of three referring expressions. The mean DICE score for A on this corpus would be $\frac{1.0+0.3+0.2}{3} = 0.5$. The summed DICE score

would be $\frac{6+3+1}{6+10+5} = \frac{10}{21} \approx 0.48$. Now, suppose system B achieved the DICE scores of $\frac{1}{5}$=0.2, $\frac{3}{10}$=0.3 and $\frac{1}{1}$=1.0 on the same data set. The average DICE score would of course be the same as for A. However, the summed DICE score for B would be lower than for A: $\frac{1+3+1}{5+10+1} = \frac{5}{16} \approx 0.31$, reflecting the fact that A achieved a perfect DICE score for a description containing six properties, while B did the same only for a description containing one property.

3.3.4 Evaluation against Multiple Gold Standards

Ideally, REG corpora should contain more than one gold standard description for each stimulus, usually from different speakers. This raises the question of how an evaluation metric should react to a system replicating one of these gold standards perfectly and getting another one completely wrong. As I argued in Section 2.5.3, if we are interested in replicating and explaining human behaviour, a REG system should be able to mimic all of the different gold standards and choose between them based on non-reference related information, such as the identity of the speaker or, for example, for how long a participant has already been involved in a data collection experiment. The simple solution commonly adopted in REG is therefore to produce a system output for each referring expression in a corpus, and to average (or sum) over all instances equally. Because all existing REG algorithms are deterministic and their parameter settings are usually not allowed to be changed during the course of an evaluation exercise, this means that they will not be able to replicate seemingly random variation in the human data. In Chapter 4, where I will be probing the in-principle capabilities of algorithms, I will adopt a different solution: I will produce the full set of all referring expressions each algorithm is capable of generating under any parameter setting, and then use Recall and Precision to check what proportion of the corpus each algorithm is able to cover and what proportion of the system-produced descriptions do not occur in the corpus.

3.3.5 Surface-Level Evaluation

Of course, human-likeness can be assessed not only at the level of semantic content but also at the level of fully realised noun phrases. The advantage of evaluating at the level of realised noun phrases is that it saves the processing step of analysing the human-produced referring expressions in a corpus and annotating them with their semantic content, which can, of course, introduce error. The disadvantage is that, in order to evaluate REG systems that are only concerned with content determination, the property sets generated by these systems need to be realised before they can be compared to human-produced noun phrases. This is again an

extra processing step and arguably one that makes it more difficult to be fair to the evaluated systems. Determining the semantic content of the human-produced noun phrases in a corpus is done according to a many-to-one (or at least many-to-few) mapping, as the meaning of at least simple noun phrases such as the descriptions in REG corpora is usually unambiguous. Realising the property sets produced by REG systems as fully-fledged noun phrases, on the other hand, is not as straightforward, as there are almost always different ways of expressing any semantic content linguistically. In other words, linguistic realisation of semantic content has to choose one solution from a one-to-many mapping. Evaluation at the realisation level therefore makes much more sense for end-to-end REG systems, whose goal it is to produce a noun phrase rather than a set of properties, than for content determination systems, which only output attribute sets. A number of such end-to-end REG systems were submitted and assessed in the 2008 and 2009 REG evaluation challenges (see Section 2.4.3). To compare the system-produced noun phrases to the human-produced ones in the TUNA Corpus, those challenges used the Levenshtein string-edit distance (Levenshtein, 1966) as well as BLEU (Papineni et al., 2001, 2002) and NIST (Doddington, 2002), two metrics usually used to evaluate the output of Machine Translation systems against human-produced translations.

3.4 Summary

In this chapter, I have discussed a number of concepts that play an important role for corpus-based work in REG. These concepts were related to the three tasks that are addressed in this thesis: corpus collection, corpus analysis and corpus-based evaluation. In the first part of the chapter, I concentrated on the different ways in which a corpus can be assembled. I argued that for the purpose of this thesis a corpus of referring expressions should

- be purposefully collected rather than extracted from found text or speech transcripts;
- consist of a set of isolated one-off referring expressions rather than chains of referring expressions from a continuous discourse;
- be based on visual stimuli that contain very simple objects in a 3D scene; and
- be collected in an online experiment.

These criteria were based on the need to control the factors that might have an impact on the content of a referring expression and to collect as much data as possible, while keeping the reference situation as natural as possible.

When analysing such a corpus of referring expressions, it is important to take into account the qualitative differences between the attributes of the objects contained in the stimuli. In Section 3.2.1, I briefly discussed the differences between type, absolute-valued attributes such as colour, gradable attributes such as size or location, and relations between objects. Also important in the context of analysing corpora are the concepts of minimality and over-specification, which I defined and discussed in Section 3.2.2. Here, I described a number of different views that have been expressed in the literature on these concepts and explained the perspective I adopt for the analysis of the corpora discussed in the following chapters. In particular, I always count the type of a referring expression when calculating its length, even though some authors have argued that it is often only included to facilitate the linguistic realisation of the referring expression as a noun phrase. I also argued that locational and relational information should be allowed to be part of a minimal description.

Finally in Section 3.3, I gave an overview of different ways to compare the output of a REG system to the human-produced descriptions in a corpus, and indicated which comparison methods I will use in the applicable sections of this thesis. It is possible to compare system output to corpus data at the level of whole referring expressions, which results in an Accuracy score of the proportion of correctly replicated descriptions. A more fine-grained approach is to use set-comparison metrics such as DICE and MASI, which take into account whether a system has produced a referring expression that is at least similar to the human gold standard description. Problems arise due to the fact that most human-produced corpora contain different descriptions for each stimulus item. Often the variation in descriptions for the same stimulus appears to be random, which means that standard REG algorithms are unable to replicate all of them under the same parameter setting. In order to assess the plausibility of an algorithm being an accurate model of human reference behaviour in Chapter 4, I therefore adopt a practice whereby I pool all descriptions the algorithm can produce under any parameter setting, and then use Recall and Precision scores to determine how well this set of descriptions matches those contained in my human-produced corpora.

Chapter 4

Corpus-Based Evaluation

The main aim of this chapter is to test three popular REG algorithms, Dale's (1989) Greedy Algorithm (GREEDY), Dale and Reiter's (1995) Incremental Algorithm (IA), and Dale and Haddock's (1991b) Relational Algorithm (RA), for their potential to be a model of human reference behaviour. To this end, I present an evaluation experiment in which I check if the three algorithms can *in principle* generate the descriptions contained in a human-produced data set. It turns out that while two of the algorithms achieve relatively good Recall results on the non-relational part of the corpus, none of the algorithms can be considered an accurate model of how people refer.

The experiment highlights a number of issues that complicate corpus-based evaluation of the human-likeness of referring expression generation. In the second part of the chapter, I discuss these issues in detail. I also take a look at how the recent evaluation challenges in REG have addressed them.

Section 4.1 introduces the data set on which the evaluation experiment presented in Section 4.2 is based. Section 4.3 contains the discussion of complicating issues in corpus-based evaluation of REG, and in Section 4.4, I examine how these issues have impacted on the recent STECs.

4.1 The Drawer Data

For the evaluation experiment described in this chapter, I use a small pre-existing corpus of human-produced referring expressions, which were drawn from a physical experimental setting consisting of four filing cabinets located in a fairly typical academic office. Each filing cabinet is four drawers high and the cabinets are positioned directly next to each other, so that the drawers form a four-by-four grid. Each drawer is labelled with a number between 1 and 16 and is coloured

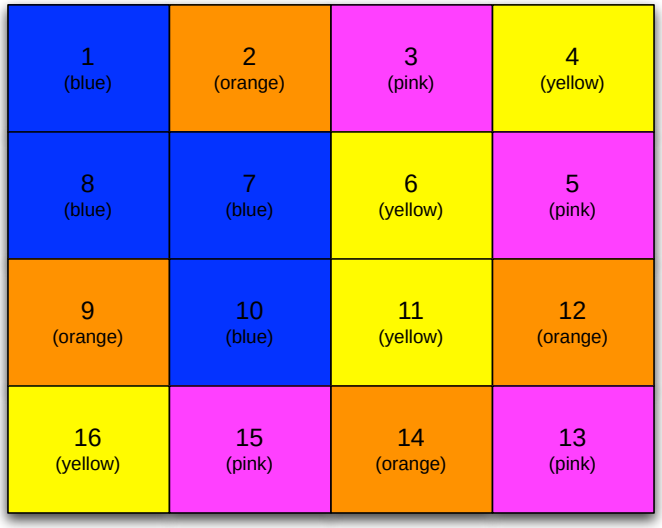

Figure 4.1: The layout of the filing cabinets

either blue, pink, yellow, or orange. There are four drawers of each colour, which are distributed randomly over the grid, as shown in Figure 4.1. I use the symbols $d_1, d_2 \ldots d_{16}$ as unique identifying labels for the 16 drawers.

Subjects were given a randomly generated number between 1 and 16, and asked to produce a description of the numbered drawer using any properties other than the number. There were 20 participants in the experiment, resulting in a total of 140 referring expressions. Here are some examples of the referring expressions produced:

(4.1) the top drawer second from the right [d_3]

(4.2) the orange drawer on the left [d_9]

(4.3) the orange drawer between two pink ones [d_{12}]

(4.4) the bottom left drawer [d_{16}]

(4.5) the drawer in the top left corner [d_1]

Each participant was asked to generate only one description in any given instance, but the same participant might have contributed several descriptions, sometimes even for the same drawer, at different points in time. The number of descriptions from each participant varies between 1 and 20. Since the selection of which drawer to describe was random, the data set does not contain an equal number

of descriptions of each drawer; it ranges from two descriptions of Drawer d_1 to 12 descriptions of Drawer d_{16}.

I did not collect this data myself, but I was the first to formally analyse and use it. As far as I am aware, the collection experiment was not based on a clear hypothesis as to the form or content of the referring expressions it would elicit. At the time of its collection no other corpus of context-free distinguishing descriptions existed — the TUNA Corpus appeared on the scene only a few years later — and so the main aim was simply to collect such a corpus in a controlled domain.

The aim of this chapter is to examine whether the REG algorithms under scrutiny are able to generate the referring expressions produced by the contributors to the corpus; since these algorithms produce distinguishing descriptions to singletons, I removed from the data set 22 descriptions which were outside the intended scope of these algorithms (essentially, those descriptions which were either ambiguous or which referred to a set of drawers in order to distinguish the intended referent). This resulted in a total of 118 distinct referring expressions, with an average of 7.375 distinct referring expressions per drawer.

As the algorithms under scrutiny here are not concerned with the final syntactic realisation of the referring expressions produced, I annotated each human-produced referring expression with its semantic content, the set of properties that were used to distinguish the target referent from the other drawers. Four direct properties used for describing the drawers can be identified in the natural data produced by the human participants. The only inherent visual attribute used is the colour of the drawer. The other distinguishing features used in this corpus are all locational in nature: the drawers' row and column in the grid, and in those cases where a drawer is situated in one of the corners, its cornerhood. I opted to annotate cornerhood separately from the row and column information because often all three are mentioned in the same referring expression, as for example in Description (4.5), despite the fact that cornerhood can be derived from the column and row information. A number of the natural descriptions also made use of the following spatial relations that hold between drawers: above, below, next to, right of, left of and between. I will discuss the representation of the attributes and their values in more detail in Section 4.2.1 below.

In Table 4.1, Count shows the number of descriptions using each property, and the percentages show the ratio of the number of descriptions using each property to the number of descriptions for drawers that possess this property (only 27 of the descriptions referred to a corner drawer). I have combined all uses of spatial relations into one row in this table, since their overall use is far below that of the other properties: 103 descriptions (87.3%) did not use a relation.

Property	Count	% (out of possible)
Row	95	79.66% (118)
Column	88	73.73% (118)
Colour	63	53.39% (118)
Corner	11	40.74% (27)
Relation	15	12.71% (118)

Table 4.1: The properties used in descriptions

The random distribution of colours results in different drawers having different characteristics regarding the discriminatory power of their properties. In particular, in a situation where colour has already been added to a description, row or column information will be less useful for an intended referent which shares colour with another drawer in the same row or column or in both, than for drawers whose colour is unique in their column and row. This might increase the chances of other properties, such as the second bit of grid information or, if possible, cornerhood, being chosen to describe these drawers and the likelihood of an over-specified referring expression.

The drawers that possess this characteristic are Drawers d_1, d_5, d_6, d_7, d_8, d_{10}, d_{11}, d_{13}, and d_{15}. Three of these, d_7, d_8 and d_{13}, share colour with a drawer in both their row and their column. With the discriminatory power of both grid properties being low after the choice of colour, the likelihood for other properties to be chosen should be even higher here. d_4 and d_{16} are special cases in that they are the only drawers with the cornerhood property that share the same colour. This reduces the discriminatory power of either of these properties once the other one has been chosen.

As we saw in Chapter 2, many algorithms in the literature aim at generating descriptions that are as short as possible, but some will under certain circumstances produce redundancy. Many authors (for example, Dale and Reiter, 1995; Arts, 2004; Engelhardt et al., 2006) have pointed out that human-produced descriptions are often over-specified, and this is borne out by the human-produced data here. However, a strong tendency towards short descriptions is evident in the data set: only 29 of the 118 descriptions (24.6%) contain redundant information. Here are a few examples:

(4.6) the yellow drawer in the third column from the left second from the top [d_6]

(4.7) the blue drawer in the top left corner [d_1]

(4.8) the orange drawer below the two yellow drawers [d_{14}]

In the first case, either the colour or column properties are redundant; in the second, colour and cornerhood, or only the grid information, would have been sufficient; and in the third, it would have been sufficient to mention one of the two yellow drawers.

One of the most obvious things about the data set is that each drawer gets described in many different ways. Even the same person may refer to the same drawer in different ways on different occasions, with the differences being semantic as well as syntactic. For example Drawer d_{11} was referred to by the same participant once as in Description (4.9) and on another occasion as in Description (4.10); and in Descriptions (4.11) and (4.12), both for Drawer d_4, the same semantic content is expressed differently at a syntactic level.

(4.9) the drawer second from the bottom and second from the right [d_{11}]

(4.10) the yellow drawer next to the orange drawer [d_{11}]

(4.11) the drawer in the top right [d_4]

(4.12) the top right drawer [d_4]

4.2 An Evaluation Experiment

Before the launch of the recent REG evaluation challenges (see Section 2.4.3) there was surprisingly little work in Natural Language Generation that compared the output of implemented systems with natural language generated by humans. Such a comparison is essential in REG if we want to assess whether the algorithms being developed can be considered models of human production of referring expressions. Conducting such an evaluation of existing algorithms against human-produced data also makes it possible to pinpoint more concretely the issues that can arise in corpus-based evaluation of REG systems, which I will return to in Section 4.3 below.

The evaluation experiment I present in this section consists of three steps:

1. the implementation of a knowledge base corresponding to the drawer domain (Section 4.2.1);
2. the re-implementation of three existing algorithms from the literature to operate in that domain (Section 4.2.2); and
3. a detailed assessment of the algorithms' performance against the set of human-produced referring expressions (Section 4.2.3).

4.2.1 Knowledge Representation

Given a Drawer d_i as target referent, the task of a REG algorithm is to produce a distinguishing description of that drawer with respect to a distractor set consisting of the other 15 drawers. I represent the drawers as one-place predicates and spatial relations between drawers as two-place predicates. Thus we have, for example, the set of properties for Drawer d_2:

{orange(d_2), row1(d_2), column2(d_2), right-of(d_2,d_1), left-of(d_2,d_3), next-to(d_2,d_1), next-to(d_2,d_3), above(d_2,d_7)}

This drawer is in the top row, so it does not have a property of the form below(d_x). The four corner drawers additionally possess the property corner(d_x).

This raises the question of what properties should be encoded explicitly, and which should be inferred. Cornerhood can be inferred from the row and column information; however, I make this property available explicitly in the knowledge base, because all of the natural descriptions that mention the target's corner position also mention its row and column; it seems plausible that this is a particularly salient property in its own right. Note in the example above that I also explicitly encode relational properties, such as left-of and right-of, which could be computed from the grid position of the drawers involved. Since none of the algorithms explored here are able to use spatial inference over knowledge base properties, I opted to 'level the playing field' by representing relations in the knowledge base explicitly. This enables a fairer comparison between human-produced and machine-produced descriptions, as the machine produced descriptions would otherwise never be able to include any relations.

A similar question of the role of inference arises with regard to the transitivity of spatial relations. For example, if d_1 is above d_9 and d_9 is above d_{16}, then it can be inferred that d_1 is transitively above d_{16}. In a more complex domain, the implementation of this kind of knowledge might play an important role in generating useful referring expressions. However, the uniformity of this domain results in this inferred knowledge about transitive relations being of little use; in fact, in most cases, the implementation of transitive inference might even result in the generation of unnatural descriptions such as (4.13) for d_{12}.

(4.13) the orange drawer (two to the) right of the blue drawer [d_{12}]

The only case in this data set where it might be logical to regard spacial relations as transitive are descriptions of the form *the orange drawer below the two yellow ones*, implying that there is a transitive below relation between d_{14} and Drawer d_6. To avoid transitivity and inferred relations, this description could

be represented directly as {orange(d_{14}), below(d_{14},d_{11}), yellow(d_{11}), below(d_{11},d_6), yellow(d_6)}, leaving it to the next step in the NLG pipeline to decide whether it should be realised as Description (4.14) or as Description (4.15).

(4.14) the orange one below the two yellow ones

(4.15) the orange one below the yellow drawer that's below another yellow one

Another aspect of the representation of relations that requires a decision is that of property hierarchies: in the drawer domain, next-to can be regarded as a generalisation of the relations left-of and right-of. The only algorithm of those I examine here that provides a mechanism for exploring a generalisation hierarchy is the Incremental Algorithm (Dale and Reiter, 1995), but it cannot handle relations; so, I take the shortcut of explicitly representing the next-to relation for every left-of and right-of relation in the knowledge base. I then implement special-case handling that ensures that, if one of these facts is used, the more general or more specific case is also deleted from the set of properties still available for the description. This is essentially a hack; however, there is clearly a need for some mechanism for handling what we might think of as equivalence classes of properties, and this is effectively a simple approach to this question.

4.2.2 The Algorithms

As we saw in Chapter 2, there is a considerable literature on the generation of referring expressions, and many papers in the area provide detailed algorithms. I focus here on the following three popular algorithms:

- The Greedy Algorithm (GREEDY) (Dale, 1989) attempts to build a minimal distinguishing description by always selecting the most discriminatory property available; it is described in detail in Section 2.2.1 and pseudocode for it is given in Algorithm 2.1.[1]

- The Relational Algorithm (RA) (Dale and Haddock, 1991a,b) uses constraint satisfaction to incorporate relational properties; see Section 2.3.1 and Algorithm 2.3 for details.

- The Incremental Algorithm (IA) (Reiter and Dale, 1992; Dale and Reiter, 1995) considers the available properties to be used in a description via a preference ordering; details can be found in Section 2.2.2 and in Algorithm 2.2.

[1]Note that in (Viethen and Dale, 2006a) we erroneously called this algorithm the Full Brevity algorithm. The differences between Full Brevity and GREEDY are explained in Section 2.2.1.

Of course, more recent algorithms are available in the literature; however, they are mostly based on the three basic algorithms listed above. As I will discuss in Section 4.2.4, the behaviour of these newer algorithms does not differ significantly from that of the basic algorithms in any way that would affect the performance on replicating the referring expressions found in the Drawer Data.

The three algorithms I examine here all follow the same pattern: the main loop runs through the prioritised list of properties of the referent, selecting a new fact to be added to the description in each iteration. In each cycle, the objects ruled out by the added fact are removed from the list of distractors and the used fact is removed from the list of available properties for the intended referent. This loop is repeated until either no more distractors are left, which means that a unique description has been constructed, or until no more properties are available for the referent, which means that no distinguishing description can be found.

The difference between the IA and GREEDY lies in the way the next property to be added to the description is chosen. GREEDY tests all available properties for their discriminatory power in each iteration and then chooses the one that rules out most distractors. The IA simply checks whether the next property in the preference order removes any distractors at all and, if it does, includes this property; otherwise it moves on to the next property. I did not implement the mechanism described in (Dale and Reiter, 1995) to handle generalisation hierarchies over the values for the different attributes, since the other two algorithms do not include such a mechanism and I therefore did not include such hierarchies in the knowledge base used for this experiment.

The RA follows the same approach as GREEDY in that it always includes the property that rules out most distractors. However, the RA is considerably more complex as it is designed to handle relations between objects. It therefore needs to be able to describe more than one drawer in the same expression, which requires keeping track of several property lists and distractor sets. This is done by adding each new object to be described to a stack and handling one distractor set for each of these objects in a constraint network. The constraints are the properties which have already been added to the description under construction. Every time a new property is added, the constraint network adjusts the distractor sets accordingly.

To avoid endless recursive descriptions, such as

(4.16) The rabbit inside a hat which contains a rabbit inside a hat ...

Dale and Haddock (1991b) propose to never use the same fact twice. In order to incorporate this solution, my implementation does not initialise a new set of properties for each object added to the object stack. Instead it removes used facts

from the knowledge base and only requests the set of properties still available for a drawer from the knowledge base when they are needed. This solution is similar to the way Krahmer et al.'s (2003) graph-based approach avoids endless recursion. In their approach, properties are represented as edges in a graph. As each edge only exists once in the domain graph, it can also only be included in the description graph once.

In an attempt to model what appear to be semi-conventionalised strategies for descriptions that people use, the IA explicitly encodes a preference ordering over the available properties. This also has the consequence of avoiding a problem that faces the other two algorithms: since GREEDY and the RA choose the most discriminatory property at each step, they have to deal with the case where several properties are of equal discriminatory power. This turns out to be a common situation in the drawer domain. Neither (Dale, 1989) nor (Dale and Haddock, 1991b) make provisions for arbitration in such cases, presumably because this did not occur in the domains they considered. It would be possible to implement a random choice; however, this would then make it hard to generate the complete set of all referring expressions these algorithms are able to produce, as is required for the assessment of their in-principle capabilities which this experiment is aimed at. It is therefore necessary to control the choice process systematically by imposing some selection strategy. I do this here by borrowing the idea of the preference ordering from the IA and using it as a tie-breaker when multiple properties are equally discriminatory. By trying all different preference orderings it is then possible to assure that all referring expressions these algorithms can produce are captured.

Type information (i.e., the fact that some d_i is a drawer) has no discriminatory power and therefore will never be chosen by any of the algorithms. Consistent with much other work in the field, I assume that the type will always be added irrespective of whether it has any discriminatory power.[2] This means that there are only four different attributes which GREEDY and the IA have to choose from: row, column, colour, and position. This results in 4! = 24 different possible preference orderings. Since some of the human-produced descriptions use all four attributes, I tested these two algorithms with all 24 preference orders.

For the assessment of RA, I added the five relations next to, left of, right of, above, and below. This results in 9! = 362,880 possible preference orderings; far too many to test. Since I am primarily interested in whether the algorithm can generate the human-produced descriptions, not in finding a 'best' preference order or in testing all preference orders, I was able to restrict the number of preference

[2]See Section 2.5.1 for a discussion of the psycholinguistic basis and some repercussions of this practice of always adding type.

orders by only considering those that begin with a permutation of the attributes contained in at least one description in the human-generated data set. Furthermore, for each of these permutations, I only had to try one randomly chosen continuation containing the remaining domain attributes. The greedy characteristic of the algorithm ensures that, if the RA will ever choose the set S of attributes given in a human description, it will definitely choose it given one of the preference orderings starting with a permutation of this set A. This resulted in 12 preference orderings incorporating the relational attributes, which I tried in addition to 24 preference orderings starting with one of the 24 permutations of the non-relational attributes which were used for GREEDY and the IA.

4.2.3 Results

Using the knowledge base described in Section 4.2.1, I applied the three algorithms to see whether they would be able to produce the referring expressions generated by the human subjects. This section discusses the extent to which the behaviour of the algorithms matched the human data.

Overview

GREEDY and the IA were used to generate 384 descriptions each, one for each of the 16 drawers using each of the 24 preference orders. The RA generated an additional 12 descriptions for each drawer using the 12 preference orders beginning with a permutation of the properties contained in the relational descriptions. This resulted in 576 descriptions from the RA. Of course, the algorithms did not generate a different referring expression for every preference order; some preference orders resulted in the same output. For example, GREEDY will produce a description containing only {column, row} for any preference order starting with those to attributes. GREEDY generated 88 distinct descriptions, and the IA generated 145 distinct descriptions. For the RA it does not really make sense to count the distinct descriptions as I did not try all possible preference orders for this algorithm. However, the particular preference orders that I used for the RA resulted in 60 distinct referring expressions.

Table 4.2 shows the number of descriptions each algorithm was tested on, the overall number of descriptions each algorithm produced, and the Recall and Precision scores for the three algorithms. Perhaps surprisingly, the Relational Algorithm does not generate any of the human-produced descriptions. Both its Recall and its Precision are therefore 0.0. I will return to discuss why this is the case below. GREEDY was able to generate 80 of the 103 non-relational descriptions

	GREEDY	IA	RA
instances in test set	103	103	118
distinct instances produced	88	145	(60)
# of distinct instances in test set	80	98	0
Recall	0.777	0.951	0.0
total instances produced	384	384	576
# of total in test set	248	236	0
Precision	0.646	0.615	0.0
F-measure	0.705	0.747	0.0

Table 4.2: Performance of the three algorithms pooled over all preference orders. GREEDY and IA were only tested on the 103 non-relational descriptions.

in the natural data set, providing a Recall of 0.777.[3] The Recall score for the IA is 0.951, generating 98 of the 103 descriptions. As these algorithms do not attempt to generate relational descriptions, the relational data is not taken into account in evaluating their performance here. Both algorithms are able to replicate all the human-produced sufficient descriptions that contain spatial relations. In addition, GREEDY —unintentionally— replicates the redundancy found in nine descriptions, and the IA produces all but five of the 29 over-specified descriptions.

Of the 384 descriptions that GREEDY produced, 248 are contained in the human-generated data set, which results in a Precision of 0.646. The IA's Precision is slightly lower at 0.615, as only 236 of its 384 descriptions were contained in the human data set. This indicates that both algorithms have to over-generate to a fairly large degree, in order to be able to achieve their high Recall scores. However, the test set used here is quite small; with more data, the likelihood of subjects producing more of the descriptions generated by the algorithms but not in the current data set would rise. At the same time, a larger data set might also contain more descriptions that the algorithms are unable to reproduce, which would in turn lower their Recall scores. This trend is indicated in the inverse relation between Recall and Precision between the two algorithms: the IA's higher Recall score is coupled with a Precision that is lower than that of GREEDY.

There are three significant points that deserve further consideration here: first, the performance of GREEDY and the IA for the individual preference orders; second, the coverage of redundant descriptions by GREEDY and the IA; and third, the inability of the RA to replicate any of the human data. In the following, I examine these three issues in more detail.

[3] In (Viethen and Dale, 2006a,b, 2007), we incorrectly reported this to be slightly higher at 0.791 (82 of 103 descriptions).

pref order	Recall GR	Recall IA	Recall mean	Precision GR	Precision IA	Precision mean	F-measure GR	F-measure IA	F-measure mean
23	**0.447**	0.398	0.422	0.126	**0.136**	**0.131**	**0.197**	**0.203**	**0.200**
4	**0.447**	0.398	0.422	0.126	**0.136**	**0.131**	**0.197**	**0.203**	**0.200**
20	**0.447**	**0.447**	**0.447**	0.126	0.126	0.126	**0.197**	0.197	0.197
5	**0.447**	**0.447**	**0.447**	0.126	0.126	0.126	**0.197**	0.197	0.197
6	**0.447**	**0.447**	**0.447**	0.126	0.126	0.126	**0.197**	0.197	0.197
19	**0.447**	**0.447**	**0.447**	0.126	0.126	0.126	**0.197**	0.197	0.197
14	0.369	0.398	0.383	0.117	0.136	0.126	0.177	**0.203**	0.190
17	0.369	0.398	0.383	0.117	0.136	0.126	0.177	**0.203**	0.190
24	0.291	0.165	0.228	0.107	0.078	0.092	0.156	0.106	0.131
21	0.291	0.165	0.228	0.107	0.078	0.092	0.156	0.106	0.131
22	0.291	0.165	0.228	0.107	0.078	0.092	0.156	0.106	0.131
18	0.282	0.165	0.223	0.097	0.078	0.087	0.144	0.106	0.125
1	0.272	0.117	0.194	0.087	0.078	0.083	0.132	0.093	0.113
3	0.272	0.117	0.194	0.087	0.078	0.083	0.132	0.093	0.113
2	0.272	0.117	0.194	0.087	0.078	0.083	0.132	0.093	0.113
12	0.272	0.117	0.194	0.087	0.078	0.083	0.132	0.093	0.113
11	0.175	0.165	0.170	0.087	0.078	0.083	0.117	0.106	0.111
10	0.175	0.165	0.170	0.087	0.078	0.083	0.117	0.106	0.111
16	0.175	0.165	0.170	0.087	0.078	0.083	0.117	0.106	0.111
13	0.262	0.117	0.189	0.078	0.078	0.078	0.120	0.093	0.107
7	0.136	0.117	0.126	0.078	0.078	0.078	0.099	0.093	0.096
9	0.136	0.117	0.126	0.078	0.078	0.078	0.099	0.093	0.096
8	0.136	0.117	0.126	0.078	0.078	0.078	0.099	0.093	0.096
15	0.136	0.117	0.126	0.078	0.078	0.078	0.099	0.093	0.096

Table 4.3: Performance of GREEDY and the IA with the individual preference orders in order of mean F-measure. The best results in each column are in bold. The exact preference orders are given in Table 4.4.

Performance of Individual Preference Orders

Table 4.3 shows the results of the 24 individual preference orders that were used for GREEDY and the IA, orderd by the mean F-scores that were achieved by using them, and Table 4.4 lists the actual orders. What Table 4.3 makes clear is that GREEDY and the IA only achieve their relatively high Recall and Precision scores because the output from the different preference orders was pooled. Neither algorithm is able to get close to these results with just one single preference order. It can therefore not be the case that each preference order represents a strategy generally applicable in a domain, as Dale and Reiter (1995) envisaged. Rather, to replicate human data from the same domain many different preference orders are necessary.

One likely reason for this is that people are different and follow different reference strategies. The drawer corpus contains too few instances per participant

23	row ≫ corner ≫ col ≫ colour
4	col ≫ corner ≫ row ≫ colour
20	row ≫ col ≫ corner ≫ colour
5	col ≫ row ≫ colour ≫ corner
6	col ≫ row ≫ corner ≫ colour
19	row ≫ col ≫ colour ≫ corner
14	corner ≫ col ≫ row ≫ colour
17	corner ≫ row ≫ col ≫ colour
24	row ≫ corner ≫ colour ≫ col
21	row ≫ colour ≫ col ≫ corner
22	row ≫ colour ≫ corner ≫ col
18	corner ≫ row ≫ colour ≫ col
1	col ≫ colour ≫ corner ≫ row
3	col ≫ corner ≫ colour ≫ row
2	col ≫ colour ≫ row ≫ corner
12	colour ≫ row ≫ corner ≫ col
11	colour ≫ row ≫ col ≫ corner
10	colour ≫ corner ≫ row ≫ col
16	corner ≫ colour ≫ row ≫ col
13	corner ≫ col ≫ colour ≫ row
7	colour ≫ col ≫ corner ≫ row
9	colour ≫ corner ≫ col ≫ row
8	colour ≫ col ≫ row ≫ corner
15	corner ≫ colour ≫ col ≫ row

Table 4.4: The 24 preference orders used for GREEDY and the IA in order of the average F-score they achieved in these two algorithms (see Table 4.3).

to systematically test this hypothesis. However, this data set contains instances where the same participant was asked to refer to the same object on different occasions, something which is not the case in the TUNA Corpus or the corpora I will describe in Chapter 5. I discussed in Section 4.1 that even the same participant sometimes referred differently to the same object on different occasions, as shown in Descriptions (4.9) and (4.10), repeated here, which were both given by the same participant.

(4.9) the drawer second from the bottom and second from the right [d_{11}]

(4.10) the yellow drawer next to the orange drawer [d_{11}]

If this participant was following the same procedure for generating referring expressions as one of our algorithms, different strategies —or preference orders— must have been at play on each occasion. This demonstrates that even allowing for a different preference order for each participant would not allow the IA or GREEDY to achieve their high overall coverage of the human-produced data in the drawer corpus. The question arises of how likely it is that speakers change

their preferences for different properties in a seemingly random way, as would be required to explain at least a large part of this data in the incremental or greedy search paradigms.

Coverage of Redundancy

Neither GREEDY nor the IA presume to be able to generate relational descriptions; however, both algorithms are able to produce each of the non-relational non-redundant descriptions from the set of natural data under at least one of the preference orderings. Both also generated several of the over-specified descriptions in the natural data set, but do not capture all of the human-generated redundancies.

GREEDY has as a primary goal the avoidance of over-specified descriptions, so it is a sign of the algorithm being consistent with its specification that it covers fewer of the over-specified expressions than the IA. On the other hand, the fact that it produces any over-specified descriptions signals that the algorithm does not quite meet its aim. The cases where GREEDY produces redundancy are those in which an entity shares at least two property-values with another entity and, after choosing one of these properties, the next property to be included is the other shared one. This situation is related to the problem, noted earlier, of what to do when two properties have the same discriminatory power. In the drawer domain, the situation arises for corner drawers with the same colour (d_4 and d_{16}), and drawers that are not in a corner but for those drawers which have a drawer of the same colour in each of the same row and column (d_7 and d_8). These observations are consistent with the point Reiter (1990a) was making when he proved that the algorithm proposed by Dale (1989) does not guarantee full brevity because of its greedy heuristic.

The IA, on the other hand, generates redundancy when an object shares at least two property-values with another object and the two shared properties are the first to be considered in the preference ordering. This is possible for corner drawers with the same colour (d_4 and d_{16}) and for drawers for which there is another drawer of the same colour in either the same row, the same column, or both (d_5, d_6, d_7, d_8, d_{10}, d_{11}, d_{13}, d_{15}).

In these terms, the Incremental Algorithm is clearly a better model of the human behaviour than GREEDY as it covers more of the over-specified descriptions. Of course, a Recall of 95.1% might be considered a good result; but we have to keep in mind that these results were achieved by trying all possible property orderings. This means that the algorithm was not simply unlucky in missing the

remaining five descriptions; rather, it is not able to replicate them *under any circumstances*. It can therefore not be a completely accurate model of the way humans produce referring expressions, even without taking into account the use of relations in descriptions. We may ask then why the algorithm does not cover all the redundancy found in the human descriptions. The five over-specified descriptions which the IA does not generate are as follows:

(4.17) {blue(d_1), row1(d_1), column1(d_1), corner(d_1)}
(*the blue drawer in the top left corner* [d_1])

(4.18) {yellow(d_4), row1(d_4), column4(d_4), corner(d_4)}
(*the yellow drawer in the top right corner* [d_4])

(4.19) {pink(d_3), row1(d_3), column3(d_3)}
(*the pink drawer in the top of the column second from the right* [d_3])

(4.20) {orange(d_{14}), row4(d_{14}), column3(d_{14})}
(*the orange drawer in the bottom, second from the right* [d_{14}])

(4.21) {orange(d_{14}), row4(d_{14}), column3(d_{14})}
(*the orange drawer in the bottom of the second column from the right* [d_{14}])

The IA stops selecting properties as soon as a distinguishing description has been constructed. For Drawer d_4, for instance, if corner was selected first, the IA might return one of the following over-specified sets of properties and then stop:

(4.22) {corner(d_4), yellow(d_4), row1(d_4)} (*the yellow drawer in the top corner*)

(4.23) {corner(d_4), yellow(d_4), column1(d_4)} (*the yellow drawer in the left corner*)

(4.24) {corner(d_4), row1(d_4), column4(d_4)} (*the drawer in the top left corner*)

The human speaker who produced Example (4.18), however, has added information beyond the point at which the target drawer was fully distinguished. The IA's failure might, in this case, be explained by our modelling of cornerhood: in the referring expressions (4.17) and (4.18), it might be the case that the noun *corner* is being added simply to provide a nominal head to the prepositional phrase in an incrementally-constructed expression of the form *the blue drawer in the top right ...*, whereas I have treated it as a distinct property that adds discriminatory power. This emphasises the important role the underlying representation plays in the generation of referring expressions: if we want to emulate what people do, then we not only need to design algorithms which mirror their behaviour, but these algorithms have to operate over the same kind of underlying representation

of knowledge. However, a different underlying representation of the properties cannot explain why the IA did not replicate the three descriptions which do not mention a corner position (Examples (4.19) to (4.21)).

Relational Descriptions

The fact that the Relational Algorithm generated none of the human-produced descriptions is quite disturbing. On closer examination, it transpires that this is due to the way RA computes discriminatory power: in this domain, the discriminatory power of relations is generally always greater than that of any other property, so the RA chooses relations first. As noted earlier, relational properties appear to be dispreferred in the human data, so the RA is already disadvantaged. The relatively poor performance of the algorithm is then compounded by its insistence on continuing to use relational properties; an absolute property will only be chosen in one of three cases:

- the currently described drawer has no unused relational properties left;
- the number of distractors has been reduced so much that the discriminatory power of all remaining relations is lower than that of the absolute property; or
- the absolute property has the same discriminatory power as the best relational one and the absolute property appears before all unused relations in the preference ordering.

In the drawer domain, none of these cases can occur until a chain of relations of at least length 3 has been added. For example, while Description (4.25) would be a typical human description of Drawer d_2, the Relational Algorithm produced the description from Example (4.26). There are no descriptions of this form in the human-produced data set. This is not surprising as they sound more like riddles someone might create to intentionally make it hard for the hearer to figure out what is meant, rather than descriptions a person might use in an actual identification task.

(4.25) the orange drawer above the blue drawer

(4.26) the drawer above the drawer above the drawer above the pink drawer

In Section 2.3, I discussed how the RA determines the discriminatory power of a relation: it takes into account not only the type of the relation (left of, above, etc.), but also the identity of the landmark. Because in the drawer domain only

ever one drawer stands in a particular spatial relation to a given other drawer, this combination of relation and landmark drawer rules out all distractors, and, therefore, has maximum discriminatory power. However, once included, the related drawer needs to be described without using its unique ID, and so a chain reaction takes place resulting in relation after relation being added. What this shows is that taking into account the landmark's identity can in many situations massively overestimate the discriminatory power that a relation to this landmark can actually add to the referring expression. In effect, this strategy assumes that the landmark is already known to the listener and will need no further identification.

In other domains, taking into account the landmark to determine the discriminatory power of a relation might not be as detrimental, for example, if it can be assumed that the landmark will not be described using another relation. However, in highly connected domains such as the drawer cabinets here, it might be more useful to disregard the landmark's identity in this calculation. Note that this would result in no relations being used by the RA for any of the drawers in this domain. While each colour, row and column, as well as the cornerhood property, rule out 12 of the 15 distractors, the relations left-of, right-of, above and below, without any information about the landmark, only rule out four distractors each (the four drawers on the edge of the grid that do not border another drawer in the relevant direction), and next-to by itself never rules out any distractors. The RA would therefore always prefer to use non-relational properties and its behaviour would then be identical to that of GREEDY. This implementation of the RA would only generate a relational description if it was unable to identify the target object otherwise, which is never the case in the drawer domain.

4.2.4 Other Approaches to Relations and Redundancy

In more recent years, new algorithms have appeared on the scene with the capability of handling relations, notably the graph-based framework and a number of relational extensions to the IA. I described these approaches in Chapter 2; here I will briefly consider their ability to address the shortcomings of the IA and the RA in terms of the generation of redundancy and relational referring expressions in the way the participants in the drawer corpus did.

Over-specification and Relations in IA Extensions

The authors of all but one of the relational extensions to the IA we saw in Section 2.3.2 specifically state that spatial relations should be tried as a last resort only. This is achieved in one of two ways: Krahmer and Theune (2002) ensure that

relations always appear at the end of the preference order after all non-relational properties; and Kelleher and Kruijff (2006) split the algorithm into two separate stages. The first stage attempts to build a non-relational description, and only if that fails does the second stage try to add relations to the referring expression under construction. In other words, both of these approaches only include a relation if no non-relational descriptions can be found. Clearly, neither of these approaches would include any relations in a domain such as the drawer cabinets where each object can be fully distinguished by its non-relational properties.

Siddharthan and Copestake (2004) use a less indiscriminate strategy to deal with relations. In their approach, the preference order is sorted by the properties' discriminating quotient (a more sophisticated version of discriminatory power, see Section 2.2.2). This also applies to relations, which can therefore appear in any position of the preference order. Unfortunately, the way this algorithm computes the discriminating quotient for relations causes it to run into the same problem as the RA in the drawer domain: a relation's discriminating quotient takes into account how many objects stand in the same relation to the target referent. As this is only one for each relation except next-to, a relation's discriminating quotient would be higher than that of all other properties, and so they would appear at the very start of the preference order. The algorithm gets called recursively for the landmark as soon as a relation is introduced, which would result in very similar chains of relations being produced as those we saw from the RA.

None of these extensions of the IA alter the termination criterion of the algorithm: as soon as a referring expression rules out all distractors, it is returned. Consequently, they would not be able to produce the instances from the drawer corpus which contain more redundant properties than the basic IA is able to include in a referring expression.

Over-specification and Relations in the Graph-Based Framework

As discussed in Section 2.2.3, the knowledge representation formalism used in the graph-based approach (Krahmer et al., 2003) makes it a promising candidate for the generation of relational descriptions. To guide the search for a referring expression, the graph-based algorithm provides the ability to make use of different weighting mechanisms when adding properties to a description. Firstly, the algorithm uses a cost function over all edges and nodes, which is used to search for the cheapest possible distinguishing graph; and secondly, a preference order over the properties and relations arbitrates between equally cheap descriptions in the same way as was done in Section 4.2.2 for GREEDY and the RA. The cost function

in combination with the preference order affords much more fine-grained control over the search strategy of the algorithm than preference orders by themselves in the IA. Relations can therefore be prioritised in a way that mimics the preferences in human data. However, trying all possible combinations of cost functions and preference orders is not feasible, and it is not clear how the necessary settings for these parameters can be determined from human data.

We showed in (Viethen et al., 2008) that it is possible to include properties redundantly using the algorithm provided in (Krahmer et al., 2003). This can be done by assigning these properties a zero cost and ensuring that they appear before all others in the preference ordering. Any number of properties can be nominated for redundant inclusion in this way, which means that referring expressions containing multiple redundancies that the IA cannot replicate are in reach of this algorithm.

However, the introduction of the cost function results in an explosion of the possible combinations of parameter settings. While this might allow the generation of a wider range of different referring expressions than is possible with any of the classic algorithms, this is likely to exacerbate the Recall–Precision trade-off we saw for GREEDY and the IA: the more different descriptions an algorithm is able to reproduce, the more likely it is to over-generate, resulting in a lower Precision score. To date, no principled method has been proposed for maximising GREEDY's, the IA's or the graph-based framework's Recall performance against a human data set that both avoids massive over-generation of descriptions not contained in the corpus and acknowledges that preference orders and cost function settings need to be changed between participants and even for the same participant.

van der Sluis and Krahmer (2005) suggest a different mechanism to allow redundancy in the graph-based algorithm: they introduce a certainty score indicating the speaker's (or the algorithm's) estimate of the likelihood that the referring expression contains enough information for the listener to identify the target referent. They then alter the algorithm's termination criterion: the certainty score of the referring expression under construction has to reach a certain threshold before it is deemed appropriate. To generate over-specified descriptions, this threshold would need to be set high enough to ensure that the algorithm keeps adding properties even after the referring expression under construction is already fully distinguishing. This mechanism, while intuitively an attractive solution, adds yet another parameter to the algorithm for which principled settings would have to be found.

4.2.5 Discussion

The evaluation experiment reported in this section was aimed at establishing the status of three popular REG algorithms in terms of their ability to reproduce human reference behaviour. The output of Dale's (1989) GREEDY, Dale and Reiter's (1995) IA, and Dale and Haddock's (1991b) RA was compared to a set of human-produced referring expressions described in Section 4.1. In my implementation, each of the algorithms takes as a parameter a preference order that prioritises the properties from which to choose for the referring expression under construction. The aim of this exercise was not to evaluate which preference order in combination with which algorithm would achieve the best results, but rather whether the algorithms would be capable *in principle* of generating each of the human-produced descriptions. To this end, the output of each algorithm was pooled over all preference orders to achieve the maximum coverage possible.

It has been argued that the different property orderings constitute independent instantiations of an algorithm and that evaluating them as one set obscures the performance of these individual instantiations (Gatt, 2007, p. 98). However, as I am interested here in whether the IA, GREEDY and the RA are *in principle* capable of producing the same descriptions as humans, I consider the property orderings as parameters that can change from person to person and from situation to situation, and therefore test all possible settings for the IA and GREEDY and all that make sense for the RA. The low performance scores of the property orderings tried by (Gatt, 2007, p. 106) for the IA are a clear indication that, if the IA is to be considered as a model of human reference behaviour, the preference ordering has to vary at least between speakers, if not also between instances for the same speaker.

The IA emerged as the most likely candidate for a model of human referring expression production as it was able to reproduce over 95% of the non-relational referring expressions in the data set. However, the analysis of the remaining 5% of descriptions establishes that the IA would under no circumstances be able to generate the redundancy found in these, which rules it out as a completely accurate model of human reference behaviour, even if we set aside relational descriptions for the moment.

The RA was added to the list of tested algorithms in order to attempt the generation of the relational descriptions in the data set. It turned out that its greedy heuristic combined with the way it computes discriminatory power for relations leads to a massive preference for relations over other properties. As a result, all descriptions generated by the RA in this domain were very cumbersome and unnatural-sounding, and none of them were contained in the human-produced

data set.

The two main observations from this experiment are:

- None of the algorithms can be considered a descriptively accurate model of human reference behaviour. From the cases where the algorithms fail, it emerges that the generation of over-specified descriptions and the generation of descriptions containing spatial relations remain major challenges for REG. The two algorithms originally proposed to address these issues are not suited, at least in the domain chosen in this experiment, to fully replicate human-produced over-specified and relational descriptions, respectively. The graph-based algorithm, while providing more flexibility in the descriptions it can produce, introduces additional parameters which would lead to more over-generation.

- No one preference order suffices to replicate the behaviour of all participants or even of one participant in a given domain. Therefore, the preference orders cannot be regarded as strategies that apply as generally as was envisaged by Dale and Reiter for the IA. If our aim is to replicate and explain human reference behaviour, a more empirical approach will be necessary, taking into account the variation both within the set of descriptions for one object and within the set of descriptions from one speaker.

4.3 Issues in the Evaluation of REG Algorithms

This section is concerned not so much with the performance of individual algorithms and whether they can mimic human data, but rather takes a meta-perspective on the task of evaluation in REG. I discuss a number of issues that arise from the experiment described in the previous section and suggest possible ways forward in tackling these issues, either in a small-scale evaluation experiment such as the one I presented in the previous section or in a community-wide shared task evaluation challenge (STEC). In Section 4.4, I will take a look at how the REG evaluation challenges of recent years have dealt with the issues I raise here.

4.3.1 Representational Choice

It is widely accepted that the input for NLG tasks is not as well-defined as it is in Natural Language Understanding (NLU) tasks. In NLU the input will always be natural language, which is processed according to the task and transformed into *a machine-usable format of some kind.* The principle decisions to be taken are whether to work on written or spoken language and whether to restrict the input

to text or speech from a certain domain. The output of an NLU task depends entirely on the nature of the problem tackled. It might be parse trees, text annotated with part of speech tags, a semantic representation of the content of the input text, or any number of other formats. In NLG, on the other hand, we are working in the other direction: there exists no consensus regarding the exact form the input provided to the system should take. The input is generally a knowledge base in *a machine-usable format of some kind*, whereas it is the desired format of the output —natural language— that is clear. As Yorick Wilks is credited with observing, Natural Language Understanding is like counting from 1 to infinity, but Natural Language Generation is like the much more perplexing task of counting from infinity to 1. The problem of determining what the generation process starts from is one of the major difficulties faced by organisers of shared task competitions in the field: the usual practice is that each researcher chooses a level of representation, and a population of that level of representation, that is appropriate to exploring the kinds of distinctions that are central to the research questions they are interested in.

As alluded to earlier, the generation of referring expressions seems to avoid this problem of lack of agreement. The task is generally conceived as one where the intended referent, and its distractors in the domain, are represented by symbolic identifiers, each of which is characterised in terms of a collection of attributes (such as colour and size) with their corresponding values (red, blue, small, large, ...). This is one of the main reasons why REG was chosen as the first NLG task for a STEC. However, this apparent agreement is, ultimately, illusory. A conception in terms of symbolic identifiers, attributes and values provides only a schema; to properly be able to compare different algorithms, we still need to have agreement on the specific attributes that are represented, and the values these attributes can take.

This is amply demonstrated by the experiment I have described in the previous section. As I employed a new domain for the purpose of this evaluation experiment, I had to first decide how to represent this domain. It turns out that this raises some interesting questions closely related to the functioning of the referring expression generation algorithms to be applied in the domain. Some of the representational primitives chosen above might seem to be uncontentious: the choice of colour, row and column in particular seem quite straightforward. However, I also explicitly represented a more controversial attribute position, which took the value corner for the four corner drawers (the attribute was not specified for the other drawers). Although this property, which we might refer to as cornerhood, can be inferred from the row and column information, I added it as an explicit property because in the drawer corpus it was often used in combination with row and column and

because it seemed plausible that having a corner position is a particularly salient property in its own right. Of course, others might not agree with this decision.

This raises the general question of what properties should be encoded explicitly, and which should be derived by means of some process of inference. In the experiment above, I also explicitly encoded relational properties, such as left-of and right-of, that could be computed from the grid location of the objects involved, and I chose not to make explicit the transitivity of spatial relations. For example, if d_1 is above d_9 and d_9 is above d_{16}, then it can be inferred that d_1 is transitively above d_{16}. Due to the uniformity of the drawer domain transitive relations could result in the generation of unnatural descriptions, such as *the orange drawer (two to the) right of the blue drawer* for d_{12}.

The decisions taken regarding the representation of cornerhood, inferrable properties in general, and transitive properties were influenced considerably by the knowledge of how the algorithms to be tested actually work. If I had only assessed different types of relational algorithms, for example, I might have implemented corners, and possibly even columns and rows, as entities that drawers are spatially related to. If all or at least some of the assessed algorithms had been able to infer properties from others, cornerhood might have been implemented only implicitly as a result of the row and column properties of the drawers. The point here is that the representational choices were guided, on the one hand, by the requirements of the algorithms; and on the other, by my intuitions about salience as derived from an examination of the data. Importantly, other researchers might have made and do make different choices based on other intuitions or observations.

From the observations above, it is evident that, in any project that focusses on the generation of referring expressions, the design of the underlying knowledge base and that of the algorithms that use that knowledge base are tightly intertwined.

The designers of a shared evaluation task or metric in this context seem to have two alternatives: either they can approach this from the point of view of assessing only the algorithms themselves; or they can assess algorithms in combination with their specific representations. In the first case, clearly the input representation needs to be agreed by all ahead of time; in the second case, each participant in the evaluation is free to choose whatever representation they consider most appropriate. The latter course is, obviously, quite unsatisfactory: it is too easy to design the knowledge base in such a way as to ensure optimal performance of the corresponding algorithm. On the other hand, the former course is awash with difficulty: even in the very simple drawer domain, there are representational choices to be made for which there is no obvious guidance and which might give advantages to some types of approaches.

4.3.2 Non-Determinism of Natural Language Choice

One very simple observation from the natural data collected in the experiment described above is that people do not always describe the same object in the same way. Not only do different people use different referring expressions for the same object, but the same person may use different expressions for the same object on different occasions. Reiter, Sripada and colleagues discuss between-speaker variation in the context of lexical choice in weather reports, where different authors attach different meanings to temporal expressions such as *by evening* (Reiter and Sripada, 2002b,a; Reiter et al., 2005). However, such speaker-dependent preferences have never, as far as I am aware, been taken into account in the development of any content selection algorithm for REG, with the notable exception of Bohnet's (2008; 2009) and Fabbrizio et al.'s (2008) entries to the recent REG evaluation competitions. Most other existing algorithms typically assume that there is one best or most-preferred referring expression for every entity they might need to describe.

Generating just one good referring expression in a given situation might be appropriate for an algorithm that is being applied in the context of a specific application. However, if our aim is to fully model human reference behaviour, we need to find a way to account for the inter- and intra-speaker variation of the kind that is found in the drawer corpus. Where referring expressions are produced as part of natural dialogic conversation, there are a number of factors we might hypothesise would play a role: the speaker's perspective or stance towards the referent, the speaker's assumptions about the hearer's knowledge, the appropriate register, and what has been said previously. However, it is hard to see how these factors can play an important role in the simple experimental setup used to generate the data discussed here: the entities are very simple, leaving little scope for notions of perspective or stance; and the expressions are constructed effectively *ab initio*, with no prior discourse to set up expectations, establish the hearer's knowledge, or support alignment. The sole purpose of the utterances is to distinguish the intended referent from its distractors. Yet despite this confined experimental setup, the participants used a variety of different property combinations to describe each drawer.

I noted earlier that one regard in which multiple different descriptions of a referent may vary is that some may be redundant where others are not. Carletta (1992), in her analysis of descriptions in the Map Task (Anderson et al., 1991), distinguishes *risky* and *cautious* behaviour in the description task: while some participants would use only the briefest references, hoping that these would do the job, others would play safe by loading their descriptions with additional informa-

tion that, in absolute terms, might make the description over-specified, but which would make it easier or less confusing to interpret. It is possible that a similar or related speaker characteristic might account for some of the variation we see here; however, it would still not provide a basis for the variation within the over-specified and sufficient subsets of the data.[4] In many cases the same participant produced different sufficient descriptions for the same object, and the same applies for varying over-specified descriptions delivered by the same participant.

Of course, it can always be argued that there is no 'null context', and a more carefully controlled and managed experiment would be required to rule out a range of possible factors that predispose speakers to particular outcomes. For example, an analysis in terms of how the speakers 'come at' the referent before deciding how to describe it might be in order: if they find the referent by scanning from the left rather than the right (which might be influenced by the ambient lighting, amongst other things), are different descriptions produced? Data from eye-tracking experiments could provide some insights here. Or perhaps the variation is due to varying personal preferences at different times and across participants; or the participants simply got bored with having to describe the same thing twice and decided to spice the task up a bit by coming up with a different description.[5]

To be able to account for variation between referring expressions, the situations in which these referring expressions were produced need to be distinguishable from one another. No computational algorithm will be able to 'guess' that, in one situation, it should generate a referring expression A for a certain target referent, and in another situation, referring expression B is to be chosen, if the information it is given about these two situations is identical. For example, if we want algorithms to take into account speaker-specific differences in referring expressions for the same target referent, the algorithm needs to be told which speaker it is to mimic in a given instance; and if we want an algorithm to mimic a boredom effect, it needs, as a minimum, access to information about how many referring expressions the same participant has already provided.

However, where seemingly random variation is at play, it might not always be possible to provide information that distinguishes between instances. Even if we simply attribute this variation to some random factor, we cannot avoid the fact that there is no single best description for an intended referent. This has a direct bearing on how we can evaluate the output of a specific algorithm that generates

[4] I use the term sufficient description for a referring expression that is neither under- nor over-specified. See Section 3.2.2 for a detailed discussion.

[5] Note that there were always several hours, and often several days, between any two instances of data collection from the same speaker.

references, no matter whether this algorithm is aimed at generating just one 'good enough' description or at fully replicating the variation that can be found in human data. Research with the latter aim might address the problem of seemingly random variation by developing non-deterministic algorithms that generate for each target referent a set of referring expressions, which are all deemed to be acceptable.

The fundamental problem that the non-determinism of natural language choice poses to corpus-based evaluation in NLG in general, and in REG in particular, consists in the large question mark it places above any gold standard corpus. However large a corpus we construct, there can be no guarantee that all correct solutions are contained in it. Thus, an algorithm's output might compare extremely badly to a human-produced corpus, simply because the perfectly acceptable expressions it generates do not happen to appear in the evaluation set; just because a particular form of reference is not contained in an evaluation corpus, we cannot be certain that it is incorrect or infelicitous.

This means that, to be a useful resource for the evaluation of REG algorithms, a corpus for REG evaluation needs to contain a large number of descriptions for each referent, as opposed to just one solution per instance. It is unlikely that such a corpus can be drawn from naturally occurring text; such a corpus would need to be constructed artificially. Nevertheless, even if we can construct such a corpus, we will always need to keep in mind that an evaluation corpus in NLG will never be truly golden: a bad evaluation result might only be due to the 'bad luck' that the perfectly viable solutions a system delivers do not occur in the corpus. The larger the corpus, however, the more confidence we can have in the evaluation results.

4.3.3 Measuring Performance

Related to the above discussion is the question of how we measure the performance of REG systems, even if we assume that we do have a gold standard corpus that contains all the referring expressions deemed acceptable for each target referent. The fact that such a corpus has to contain many referring expressions per referent makes the comparison between system output and gold standard references non-trivial. There is no simple yes–no question to be answered of the form *did the system produce the same expression as the human participant?* Instead we need to answer questions such as *did the system produce an expression that was the same as one of those produced by the human participants?* or *how many expressions did the system produce that were identical to those the human participants produced?*

The problems that the inherent non-determinism of natural language choice

causes for evaluation are not unique to NLG: recent evaluation exercises in statistical machine translation and document summarisation, both tasks that like NLG have natural language as their output, have faced the problem of multiple gold standards (see Papineni et al., 2002; Nenkova and Passonneau, 2004, respectively). In both these fields evaluation metrics have been developed that compare one output text to several gold standard reference texts at the same time. However, it is not obvious that a fine-grained task such as referring expression generation can be evaluated in the same way. It might be appropriate to give credit to longer text samples, such as the ones at stake in summarisation and machine translation, for bearing resemblance in different parts to different gold standard texts; but it is not clear whether one, comparably short, referring expression can be evaluated regarding its similarity to a number of different human-produced references. A referring expression is more likely to be 'good' if it is identical, or similar, to *one* of the gold standard expressions than if it incorporates bits of all of them.

In the experiment above, I instead let the candidate algorithms produce a set of referring expressions for each target referent and then compared this set to the set for the same target referent contained in the human-produced corpus. The rationale underlying this approach was to assess the algorithms' status as descriptive models of human reference behaviour. The primary aim of the experiment was to find out whether one of the existing algorithms might be able to generate all of the human-produced referring expressions in the corpus.

To see exactly how this was done requires some understanding of how the IA works. The IA explicitly encodes a preference ordering over the properties available to be used in descriptions: so, for example, in describing an object in a physical scene, it is very common to first use the colour of the object, even if this property ultimately does not add anything to the discrimination provided by the other parts of the referring expression, so colour might appear very early in the preference order. In my implementations of GREEDY and the RA, I also included a preference ordering in order to force a choice in those cases where two properties rule out the same number of distractors. In the case of the IA, the properties are considered in the order prescribed by the preference list and a particular property is used in the referring expression if it provides some discriminatory power, otherwise it is skipped. The use of an explicit preference ordering over properties was introduced by Dale and Reiter (1995) as a way to facilitate porting the algorithm to new domains, since all one needs to do is define an appropriate ordering over the properties available in the domain. However, even within a single domain, one can of course vary the preference ordering to achieve different effects. In this way, the orderings can just as well be interpreted as personal preferences of different speakers

or reflect any number of other environmental factors, such as different degrees of salience accorded to different properties by different individuals at different times. It was by means of manipulation of the preference ordering that it was possible to generate more than one referring expression for each target referent.

Of course, such an approach is also likely to produce a large collection of referring expressions that are not evidenced in the data. If the aim is to evaluate algorithms' ability to explain human reference behaviour, they should be assessed not only by their ability to reproduce the descriptions in a corpus (Recall), but also by the number of descriptions they generate that are not contained in the corpus (Precision), to measure the balance between under-generation and over-generation. Recall and Precision can then be combined into the conventional F-measure.

Of course, not every evaluation exercise in REG is likely to be aimed at an algorithm's ability to model the human data as closely as possible. In fact, much research in REG has been aimed at building algorithms that can generate one referring expression that can be used in a given situation. Some approaches attempt to generate what is deemed to be the one best referring expression, but as we can see, even in a corpus as small as the Drawer Data, in many cases there might not exist one best referring expression. However, application-oriented REG algorithms need necessarily be deterministic in the sense that they have to decide on one referring expression to be used by the application in which they are employed. It would be highly impractical for a direction-giving system to recite a list of all the possible descriptions of a certain landmark it wants to use to indicate to a user where to turn off the main street, although even for such a system it might be advantageous to generate a number of different options, in case the first option realised fails at letting the listener identify the target referent.

The best way to use corpora in assessing such an algorithm's ability to produce natural-sounding felicitous referring expressions should be to accept only one referring expression per target referent and then check if this referring expression is contained in the human-produced set for this referent. In other words, to assess application-oriented algorithms, only a Precision-based metric should be used for comparison against a human-produced corpus.

It is worth repeating the cautionary note from Section 4.3.2: this evaluation approach, just as any corpus-based evaluation exercise in NLG, still suffers from the problem that we can never be sure how comprehensive the gold standard data set is in the first place. This impacts mostly on the reliability of Precision-based evaluation measures, because these penalise a system for generating instances that are not found in the gold standard. However even in the face of this uncertainty,

we can be sure of two things:

- a REG algorithm which is unable to generate all referring expressions contained in a human-produced corpus under all possible settings of its parameters cannot be a descriptively adequate model of human reference production; and
- a REG algorithm that only generates referring expressions that are also contained in a human-produced corpus can safely be used in an application where natural-sounding descriptions are needed.

What I have argued here is that asking the question *Does the algorithm generate the correct referring expression?* does not make sense when there are multiple possible correct answers. Instead we can ask one of two questions, depending on the research goal we are pursuing:

1. Does this algorithm generate a referring expression that a person would use?
2. Can this algorithm generate all referring expressions that human speakers have contributed to a corpus and only those?

4.3.4 Domain Specificity

Early algorithms for the generation of referring expressions, such as those evaluated in the experiment described above, were very rarely formally tested or even developed on the basis of a solid data set of human descriptions of objects. The closest this work came to an evaluation was to sketch a few worked examples, typically from a simple toy domain. These mini-domains usually consist of not more than a few objects: a couple of bowls, cups and tables, or a few animals of different types, sizes and colours.

Some more recent approaches use production experiments involving human participants for the development or evaluation of their algorithms. The algorithm presented by Funakoshi et al. (2004) is based on the analysis of human data obtained from experiments in a handcrafted abstract domain of uniform objects. van der Sluis and Krahmer (2004a) and van der Sluis and Krahmer (2004b) draw on production experiments to verify assumptions made by the algorithm they describe in (Krahmer and van der Sluis, 2003). Gatt (2006) describes the only research I know of conducted before the launch of the more recent Generation Challenges in the area of referring expression generation where algorithm performance is directly compared to human performance. However, it is not the

referring expressions themselves, but the underlying clustering of objects that is at the centre of interest in this work.

In all cases, the domains on which the assessment is based are handcrafted and rather artificial. Although cautious claims are made regarding the portability of the algorithms to other domains, these are never tested. Ultimately, most algorithms for the generation of referring expressions are designed with a certain domain in mind; if they are systematically tested at all, then it is on this one domain and against data from experiments in the same domain.

The surprisingly bad results of the Relational Algorithm in the evaluation experiment discussed above show that this domain specificity of algorithms for the generation of referring expressions makes it extremely hard to compare existing approaches. While the RA might perform well in the toy domain used for the worked examples in (Dale and Haddock, 1991b), it never had a chance in the still relatively simple domain of drawer cabinets. With hindsight, it becomes obvious that the toy domain used in that work is not well-suited for testing the ability of the algorithm to choose between relational and non-relational attributes in the way people do. The only non-relational property in the domain used in (Dale and Haddock, 1991b) is the type of the objects, which is added in all cases to provide a head for the nominal expressions produced. Consequently, the only way to make a distinction between objects of the same type, for a human speaker or for the algorithm, is to use spatial relations. At the same time, the example domain in (Dale and Haddock, 1991b) is so small that one relation always suffices to distinguish the referent from the other objects, thereby avoiding the long chains of relations we saw the RA produce in the drawer domain.

The problem of an implicit domain specificity in approaches to referring expression generation is one main reason to argue for a shared test domain. Researchers developing a new algorithm, or hoping to improve an existing algorithm, are only able to verify their advances if they can compare old and new systems in a controlled test environment.

However, this issue also points to the implausibility of 'blind development' for an evaluation competition in REG where the test domain is only revealed after development is concluded. This is common practice in other shared evaluation task communities; but the fundamental differences between Natural Language Understanding and Natural Language Generation mean that we are still far from being able to develop any kind of NLG system that is portable to a new domain without considerable effort.

4.3.5 Interim Summary

In this section, I have discussed four issues that need to be tackled in corpus-based evaluation in general, and in particular in STECs on REG. Some of these issues are applicable to NLG more widely, while others are specific to REG.

The design of REG algorithms is usually tightly intertwined with knowledge representation formalisms and often their performance depends heavily on details of the underlying representation of the entities and their properties in the referential domain. One way of addressing this problem would be to evaluate each algorithm in conjunction with a knowledge representation that allows it to produce its best possible results. However, if we are looking for the best algorithm for a task on a given knowledge representation or if we are for some other reason interested in the differences only between the algorithms rather than algorithm–knowledge base pairs, this is not possible. For a community-wide STEC this means that the way objects and their properties will be represented in the test phase needs to be known in advance to all participants.

The non-determinism of natural language choice poses one of the hardest problems for corpus-based evaluation in REG. Every object or entity can be described in many different ways and each of these descriptions might be equally acceptable. Therefore, it would be unfair to collect one description for each target referent in a domain and to then expect an algorithm to reproduce exactly this set of descriptions, one for each referent, when a large number of different descriptions might be equally acceptable. What is needed are corpora that contain as many descriptions as possible for each target referent. An algorithm can then be evaluated against this corpus in one of two ways: either its usefulness in an application can be assessed by checking its precision in generating only referring expressions also contained in the corpus, or its descriptive adequacy as a model of human reference behaviour can be assessed by checking how well the set of all referring expressions that it can produce under any parameter setting matches the set of referring expressions contained in the test corpus. Of course, if our ultimate aim is to explain the variation that occurs in human-produced referring expressions, it is necessary to attempt to find features by which the situations spawning different descriptions can be distinguished in the underlying knowledge representation. If this is possible, algorithms need to take these features into account and attempt to replicate the natural variation.

One issue that should be addressed specifically by organisers of STECs is the domain specificity of most existing REG algorithms. Most existing REG algorithms are highly domain specific due to the fact that they were designed with a usually

relatively small example or evaluation domain in mind. This domain specificity can lead to unintended results once the algorithm is tested in a new domain, as we saw for the RA in the experiment above. For the organisation of a STEC this means that the domain has to be made known before the development phase, as it is unlikely that 'blind development' and then testing on entirely unknown domains, as is done in other fields, will work in REG.

4.4 The Referring Expression Generation Challenges

The First NLG Shared Task and Evaluation Challenge on Attribute Selection for Referring Expressions Generation (ASGRE) was held in 2007 as a pilot event, both to gauge the interest within the NLG community in a STEC and to assess whether issues such as the ones discussed in the previous section would prove to be unsurmountable obstacles or whether a STEC could be turned into an opportunity to overcome these difficulties. The exercise was deemed a success and repeated in 2008 and 2009.[6]

4.4.1 The Problem with Representational Choice

In Section 4.3.1, I noted that the functioning of most REG algorithms reported in the literature is tightly intertwined with the structure and content of the underlying knowledge base on which they operate. This means that in a STEC a decision must be made between on the one hand letting participants design their own knowledge base and then assessing the algorithms in conjunction with these, and on the other hand providing the knowledge base and forcing the participants to gear their algorithm towards it. In the evaluation competitions based on both the TUNA and the GREC data, the organisers went for the second option. A more detailed examination of the TUNA Corpus reveals that the designers of the corpus had to face decisions regarding the representation of different properties that are not dissimilar from the decisions that I had to take for the experiment described in Section 4.2, and their choices seem similarly idiosyncratic.

For example, in the people domain, the fact that someone has a white beard is expressed using two attribute–value pairs: ⟨has-beard:TRUE⟩, ⟨hair-colour:white⟩. If the same person is also bald, he will also have the attribute–value pair ⟨has-hair:FALSE⟩. So, a person can be tagged both as having white hair colour and not having hair. As long as the beard is also chosen to be mentioned in a referring expression, that might not constitute a big problem, although this would presuppose

[6] See Section 2.4.3 for an overview of the setup of the REG challenges.

that the realisation step following content selection is able to interpret hair-colour as beard colour for bald people. However, it might easily yield strange results: if ⟨has-hair:FALSE⟩ and ⟨hair-colour, white⟩ are included in a referring expression, but ⟨has-beard:TRUE⟩ is not, the following description could be produced:

(4.27) the man of white hair colour without hair

A second example where knowledge representation might affect the performance of algorithms is the representation of spatial information. For the experiment reported above, I represented locational information in the drawer domain in several ways: as absolute row and column co-ordinates, as reciprocal spatial relations between the objects in the domain at different levels (above, below, left-of, right-of and next-to), and as position in the domain (cornerhood). In the TUNA annotation, on the other hand, locational information is represented only in form of the x/y-coordinates of each object in the display grid. This seems to make sense for this data set, as the human data contains no spatial relations between objects. However, representing spatial relations in one domain, but not another, means that the same algorithm will give very different output in the two domains: not because of a clever mechanism that lets it realise that spatial relations make less sense in one domain, but simply because the design of the knowledge base forces it to. Of course, it can be argued that it might not be the process of referring expression generation that results in people using spatial relations more in one domain than another, but precisely their different underlying representations of the two domains; and it is also likely that this representation differs for different people, not only for different domains. However, this only underlines how critical the interplay between an algorithm and the underlying knowledge representation is.

Whatever the reason for choosing a particular knowledge representation, it seems to be fairer to let all participating systems in a STEC operate on the same one, rather than allowing different ones. This allows an assessment of different algorithms independently of the processes involved in developing representations of domains, and it is what the organisers of the REG STECs did by imposing the TUNA annotation as the input representation to be used by participating systems.

4.4.2 The Problem with Non-Determinism

In Section 4.3.2, I concluded that, to address the inherent non-determinism of natural language, an evaluation corpus for REG needs to contain as many instances for each referential situation as possible. Algorithms can then be evaluated either by Precision alone, if they produce only one description for each target referent, or

by Recall and Precision, if they attempt to fully model human reference behaviour and produce all descriptions they deem acceptable for each referent.

In the TUNA Corpus a trial is defined as the specific set of objects contained in a scene. The location of the objects in the scene and in relation to each other is not taken into account in this definition. So, two scenes containing the same set of objects are considered to be the same trial, even if the objects appear in different locations in the two scenes. There are seven different trials with a singular referent in the furniture domain of TUNA and six in the people domain. However, there is no indication of how many different spatial arrangements were used for the different instances of a trial. This applies even to the +LOC condition of the corpus, where participants were told that location was a useful feature.

The singular portion of the original TUNA Corpus contains 60 different instances for each trial, one from each participant, so the corpus seemingly fulfils the criterion of several instances per trial. However, this definition of a trial is not equivalent to my definition of a referential scenario or even a context set due to the omission of locational information. It is likely that this was done under the assumption that location would not have much of an impact on the content of the referring expressions in the two domains of the corpus. This seems to be a false assumption to make, as the content of a referring expression in TUNA's domains might well be dependent on the position of the target referent within the scene. For example, it is likely that x- and y-coordinates would get used more for targets in extreme corner positions; and the use of locational information might impact on which other properties are included, as less inherent properties are necessary for identification if the target referent's location is mentioned as well.

A domain in the sense of the TUNA Corpus is what I call a context set, a collection of objects with all their properties, including their locations. For each context set, or domain in the sense of the TUNA Corpus, the original corpus contains only one instance. For the 2008 competition, new test data was collected under the same circumstances as the original corpus to ensure that two instances per context set were available. This methodology was also used in the 2009 competition. However, the evaluation procedure expected the candidate systems to produce one output for each of these two instances although the underlying referential scenario was identical. If the two instances of one scenario contained different referring expressions, the candidate systems had little chance of replicating both correctly. Some systems attempted to use the participant ID to distinguish referential scenarios (Bohnet, 2008, 2009; Fabbrizio et al., 2008), but this information was not always provided in the test data.

Of course, there might always be influencing factors that are not taken into ac-

count in an experiment. For instance, it could be that the age of the participant or the time of day have a large impact on the content of referring expressions. However, in the TUNA Corpus there was one intentional experimental factor that is not annotated in the data set: in the original data collection experiment, participants were led to believe that they were communicating with a language understanding system which would attempt to remove the intended referent from the scene. After a participant had typed a referring expression, one object (or in the plural trials, two objects) was removed; however, the choice was independent of whether the referring expression was accurate or not. It is likely that this type of feedback impacts people's reference behaviour. For example, after getting negative feedback, a participant might choose to produce more over-specified referring expressions. In the corpus annotation there is no information about whether the participant had received negative feedback in the previous or any preceding trial.

In acknowledgement of the problems inherent in evaluation against a corpus, the REG challenges also included corpus-independent evaluation. The candidate systems' output was automatically tested for minimality and for uniqueness — whether the target referent was fully distinguished from all distractors— and a task-based experiment assessed the speed and accuracy with which human participants were able to identify the target referents based on the system-produced descriptions. While these measures capture different views on optimality, their use does not alleviate the problem that human-likeness is difficult to assess on the basis of only one or two instances per context set.

4.4.3 The Problem with Measuring Performance

Each participating system in the REG challenges had to submit exactly one solution for each instance in the corpus. It was assessed on how well it replicated the referring expressions in these instances using the DICE set similarity co-efficient and, in 2008, MASI and a number of string based comparison metrics.[7] Given the discussion in Section 4.3.3, this is not the ideal way to evaluate human-likeness. Firstly, the systems were not allowed to switch parameter settings between instances, not even between those that were based on the same context sets in the 2008 and 2009 campaigns. The rationale of introducing a second instance for each context set was to give the systems a better chance of getting it right in face of the fact that people do not always produce the same reference in the same referential scenario. However, by not allowing the systems to try different strategies for those instances, the opposite effect can result: if the two human referring expressions

[7] See Sections 3.3 and 2.4.3 for more detail on the metrics used in the REG STECs.

for one context set were different, the systems were forced to get at least one of them wrong, because they were forced to generate the same referring expression for both. If they were the same, on the other hand, the systems had a small chance of getting them both right and a much larger chance of getting them both wrong. A slightly fairer way to evaluate systems that are only allowed to generate one referring expression for a given context set against a data set containing two instances per context set would be to only count the higher of the two scores that a system achieved on these two instances. Secondly, even with two instances per context set, the chances of producing a referring expression that is not contained in the corpus but is perfectly acceptable and might be produced by a human speaker are still very high.

4.4.4 The Problem with Domain Specificity

The 2007 and 2008 REG challenges established the furniture and people domains of the TUNA Corpus as something of the standard domains for content selection for referring expressions. The organisers successfully addressed the problem that domain-specificity of algorithms can pose for comparative evaluation by making the domains and development data available well in advance of the test round. This meant that algorithm developers had sufficient time to adapt their systems to the test domain.

However, in 2009, pure content selection was dropped from the evaluation campaign, and in 2010 the TUNA Corpus was dropped entirely as a test domain with the argument that further evaluation challenges in the same domain would not attract enough community interest to be justified. This underlines the point made above that systems need to be tested in more than just one domain in order to keep pushing the boundaries of algorithm development. For example, it would be interesting to see how many of the systems developed for the TUNA domain could be successfully be adapted to a domain in which spatial relations are frequently used to identify target referents.

4.5 Conclusions

In the first half of this chapter, I presented an experiment in which three popular algorithms for the generation of referring expressions were evaluated against a small corpus of descriptions of drawers in a grid of filing cabinets. The immediate conclusions from this experiment are:

- Despite achieving relatively high recall scores, GREEDY and the IA are not

descriptively adequate models of human behaviour. Both are able to replicate some of the over-specified descriptions in the Drawer Corpus, but there are instances that even the IA would never be able to generate under any parameter setting. Furthermore, both algorithms have to over-generate to a large degree in order to get their high Recall scores, although this might be due to the small size of the evaluation corpus.

- The generation of referring expressions that contain spatial relations emerges as one of the major challenges still to be faced by REG. In the drawer domain, the RA was proven to be incapable of producing descriptions that bear any resemblance to human-produced descriptions. Existing approaches to extend the IA with relation handling capabilities only use relations when there is no other alternative. As the relational descriptions contained in the Drawer Corpus prove, this is not what people do.

- Much of the variation in the data is due to participant-specific and even seemingly random variation. The classic algorithms in REG and the vast majority of recent approaches do not take such variation into account.

Based on the experiment in the first half of the chapter, I examined in more detail the fundamental problems that have to be addressed when REG algorithms are to be evaluated for human-likeness against a corpus of human-produced referring expressions. The conclusions I draw from this discussion are:

- The design of most REG algorithms is entwined with the nature of the underlying knowledge representation of the domains on which they operate. When evaluating different algorithms against the same data set, it is important to design the input representation in a way that does not give an advantage to one algorithm over the others. In a competitive evaluation exercise this means that the underlying knowledge representation needs to be available to all participants in advance of the development phase.

- To deal with the non-determinism of natural language choice, we require large corpora that contain many instances of referring expressions from different participants for each stimulus item. The more instances we can obtain, the more confident we can be that an evaluation based on the corpus is fair. However, no matter how large a corpus is, it remains important to keep in mind that no corpus is ever likely to be complete. No existing corpus of referring expressions is adequate in this respect.

- If we are truly interested in evaluating the potential of a REG algorithm to explain human reference behaviour, it needs to aim to replicate all referring expressions in a corpus, even if the differences between descriptions for the same stimulus item cannot be explained in a principled way. A Recall score can give us an indication of the algorithm's performance in this regard. If we are confident that the corpus is large enough to contain most of the referring expressions any person would be likely to use, a Precision score can be used to indicate how much the algorithm has to over-generate in order to achieve its Recall.

- If we are only interested in evaluating an algorithm's ability to produce one acceptable referring expression for each stimulus item (e.g., in an application-oriented context), a Precision score can be used to assess how many of the referring expressions generated by the algorithm are contained in the set of human-produced descriptions for each item.

Chapter 5

Collection and Analysis of Two REG Corpora

In this chapter, I present data collection experiments for two new corpora of referring expressions, followed by analyses of these corpora which focus on the use of spatial relations. Both corpora are based on visual stimuli of simple 3D scenes containing a small number of geometric objects. They were collected by directing participants to a website, rather than in a face-to-face setting, which allowed the collection of a very large number of samples. The first corpus, GRE3D3[1], which can be regarded as something of a pilot study for the second collection experiment, contains 630 referring expressions, which is comparable to the 780 referring expressions for singular target referents in the TUNA Corpus. The second corpus, GRE3D7[2], is at 4480 referring expressions by far the largest existing collection of context-independent distinguishing descriptions. With these corpora, I attempt to overcome the limitations of existing corpora that I have discussed in the previous chapters. The main findings from the analyses that I present in this chapter are that people regularly use spatial relations between objects even when they are not necessary, and that the use of spatial relations is impacted to a large degree by the individual preferences of the speaker and by the visual salience of the potential landmark object.

Section 5.1 outlines the aims of the collection experiments, including general reasons for collecting the new corpora and more specific research questions that they target. Sections 5.2 and 5.3 describe the collection and analysis of the first corpus. The following two sections (5.4 and 5.5) do the same for the second corpus.

[1]GRE3D3 stands for '**G**eneration of **R**eferring **E**xpressions in **3D** scenes with **3** Objects'. The corpus is available online at `http://www.science.mq.edu.au/~jviethen`.

[2]GRE3D7 stands for '**G**eneration of **R**eferring **E**xpressions in **3D** scenes with **7** Objects'. It is available at the same address.

In Section 5.6, I compare the variation found in the two corpora in terms of the content patterns of the referring expressions they contain. Finally, in Section 5.7, I provide a general discussion of the results of the analyses in this chapter.

5.1 Aim of the Corpus Collections

In Section 2.4.1, I discussed existing corpora that contain referring expressions, and we saw another such corpus in Chapter 4. For different reasons, none of these corpora are ideal for the study of the issues I want to explore. The requirements for my corpora that arise from the experiments and discussions of the previous chapters are:

1. *Visual stimuli*: in order to gain control over the set of entities that people take into account as they build a distinguishing description, I require the stimuli to be visually available.

2. *Context-free reference*: in order to be able to discern the factors that play a role in how people refer to visually available stimuli, I abstract away from any further linguistic discourse.

3. *Many instances*: in order for my corpora to be suited to the study of human reference behaviour in the face of the non-deterministic nature of natural language choice, they need to contain as many referring expressions as possible

   - for each referential scenario, and
   - from each participant.

4. *Spatial relations*: in order to be able to study the use of spatial relations between objects, the stimulus scenes have to be designed such that the use of spatial relations is encouraged but not necessary. The spatial aspects of the stimuli should be as natural as possible, which leads me to use 3-dimensional scenes rather than flat displays of objects.

The MapTask, COCONUT and iMap corpora are collections of dialogues with annotated referring expressions. These corpora are helpful for the exploration of the impact that contextual discourse factors have on the form and content of referring expressions; however, I aim to concentrate on the generation of context-free distinguishing descriptions based only on visually stimuli, which allow me to control for discourse and other external factors as much as possible. The TUNA Corpus and the Drawer Data come closest to fulfilling my requirements. However,

the Drawer Data is a rather small corpus and very unbalanced in terms of the number of descriptions it contains for each of the referential scenarios. The TUNA Corpus, while being much larger than the Drawer Data, only contains one instance for each referential scenario, if spatial arrangement is taken into account. In Section 4.3.2, I argued that the non-determinism of natural language choice demands that, in order to enable the study of human reference behaviour and evaluation of REG algorithms for human-likeness, a corpus needs to contain as many referring expressions for each referential scenario as possible. The main aim of the work presented in this chapter is, therefore, to collect large corpora that contain many referring expressions for each scenario.

The analysis of the evaluation experiment presented in the previous chapter showed that spatial relations in referring expressions remain a challenge for REG algorithms, in particular in cases where other properties can be used to identify the target referent. Most researchers who have proposed algorithms, or extensions to algorithms, which are capable of handling relations between objects insist that this option should only be taken if no relation-free distinguishing description can be found (Krahmer and Theune, 2002; van der Sluis and Krahmer, 2005; Kelleher and Kruijff, 2006). This strategy is based on psycholinguistic evidence claiming that inherent properties are easier to perceive and process than relations between objects. However, none of these algorithms have been tested against human-produced data, and the instances of relational referring expressions in the Drawer Data, where relations were never necessary to identify the target referent, suggest that at least some speakers do not follow this strategy. In order to facilitate further investigation of this issue, the corpora I present in this chapter are designed in a way that makes the use of spatial relations possible, but never necessary. The scenes presented to the human participants as visual stimuli vary on systematically chosen dimensions, making it possible to explore different factors that might influence the use of relations. In particular, I am interested in investigating how the likelihood of the spatial relation between the target referent and a potential landmark being mentioned is impacted by the visual salience of the landmark object and by the ease with which the target referent can be described using inherent properties only. Sections 5.2.1 and 5.4.1, in which I describe the stimulus design for the two corpora, will elaborate the specific ways in which visual salience was manipulated.

Note that the spatial relations I investigate here are only a subset of the overall locational information about an object. The main focus is on relations to other objects, not on more global locational information such as *in the left*, which is essentially gradable or vague (Gatt et al., 2007). I therefore attempted to design especially the stimuli for the second corpus in a way that would discourage the

use of non-relational location information. Where it nonetheless occurred in the resulting corpora, I report its use, but treat it separately from relational location information.

5.2 Collecting GRE3D3

5.2.1 Stimulus Design

In order to keep the stimulus scenes in GRE3D3 as simple as possible, I minimised the number of objects in each scene. To explore even the most basic hypotheses with respect to the use of relational expressions, at least three objects per scene were required. One of these is the **target** referent, which the participant has to describe in such a way as to distinguish it from the other two objects in the scene. Although the scenes are designed such that spatial relations are never *necessary* to distinguish the target, they are set up so that one of the two non-target objects is clearly closer to the target. I call this object the (potential) **landmark**; the third object in the scene I call the **distractor**.

I only used very simple shapes in the stimulus scenes, to keep control over the different attributes which participants were likely to use in their referring expressions; each object was either a cube or a ball. The objects varied in two further attributes: colour (either green, blue, yellow, or red); and size (either large or small). To make the spatial layout of the scenes look as natural as possible, they were drawn using the 3D drawing program Google SketchUp[3].

Research Questions

The design of the stimulus scenes was based on the exploration of three initial research questions:

1. Is the decision to use a spatial relation impacted by the similarity between target and landmark?

2. Is the decision to use a spatial relation impacted by the similarity between landmark and distractor?

3. Is the decision to use a spatial relation impacted by the length of the inherent MD[4] for the target?

[3] `http://sketchup.google.com`

[4] The shortest possible description which does not include location or any relations. See Section 3.2.2 for a definition and discussion of this term.

Questions 1 and **2** are based on the hypothesis that the visual salience of the landmark impacts the likelihood of it being included in a referring expression: the more an object catches the speaker's eye, the more likely he is to mention it in a referring expression.[5] The literature on visual salience suggests that visual salience is based on attribute difference (for an overview, see Yantis and Egeth, 1999; Caduff and Timpf, 2008; Pashler, 1998, Chs. 1 and 2). A rare attribute value makes an object stand out from the objects around it and thereby catches the onlooker's eye. For example, a blue object among green ones is visually salient and so is a right-slanting stroke among left-slanting ones. Based on this definition for visual salience, I arrive at **Hypothesis 1**.

> **Hypothesis 1:** The less similar the landmark is to the other two objects in its visual properties, the more likely it is to be mentioned.

In the scene design, I attempted to keep other factors, such as the distance between objects and occlusion of objects, as constant as possible, so that an object's visual salience would mainly be influenced by the number of properties it shares with the other two objects. This take on visual salience is related to the concept of an inherent MD mentioned in **Question 3**: the fewer properties the target shares with the objects surrounding it (i.e., the less similar it is to the other two objects), the shorter its inherent MD. In other words: an object that can be distinguished from its distractors by one attribute only is more visually salient than one for which all attributes have to be listed. At the same time, participants might be more inclined to choose a very short inherent MD over a relational description; but if the inherent MD is relatively long, a relational description might be more likely. The second hypothesis I aim to test is therefore **Hypothesis 2**.

> **Hypothesis 2:** The human participants are less likely to use the relation to the landmark, if the target can be described with a very short non-relational description.

Note that for the analysis of reference behaviour in the simple scenes used in the corpora in this chapter, it is sufficient to adopt this rough notion of visual salience. However, it is not clear how it would carry across to more complex scenes with more properties and many more objects. For example, it is not obvious whether an object that shares its type with all objects in the scene, but is unique in all its other properties, would be more visually salient than an object that has a unique type but shares each of its other properties with one object each.

Tests for **Hypotheses 1** and **2** are reported on pages 135 to 138.

[5] From a more listener-oriented perspective, the inverse is also true: the more an object catches the *listener's* eye, the more useful it will be to help him find and identify the target object.

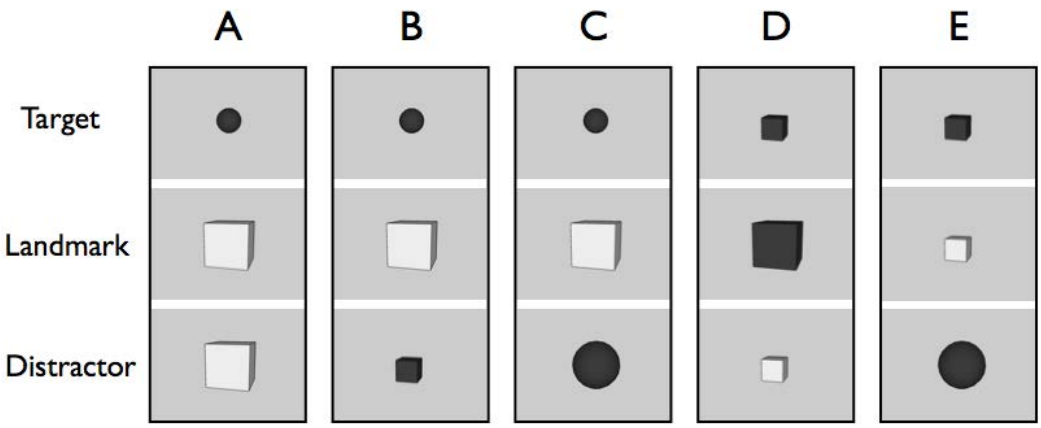

Figure 5.1: The schemata which form the basis for the GRE3D3 stimulus scenes.

Scene Schemata

With these three research questions in mind, I created five **schemata** (see Figure 5.1) as a basis for the final stimulus scenes. A schema determines the type and size of each object in the scenes that are based on it, and defines which objects share colour. So, for example, in scenes based on Schema C, the target is a small ball; the landmark is a large cube; the landmark has a different colour from the target; and the distractor is a large ball sharing its colour with the target.

I chose to make all landmark objects cubes, because it might look unnatural to see an object balanced on top of a perfectly spherical ball. I also excluded variation in the size of the target object from the analysis, making all target objects small; this avoids situations where a smaller landmark might be obscured by the target placed directly in front of it.

In Schemata A, B and C, target and landmark share no properties; and in Schemata D and E, they share two properties, which is as similar as they can be without being identical. This allows the investigation of target–landmark similarity, which is at stake in **Question 1**.

In Schema A, landmark and distractor are identical; in Schema E, they are completely distinct from each other; and in Schemata B, C and D, they share only one property (their size in Schema C, and their type in Schemata B and D). This allows the investigation of landmark–distractor similarity, which is at stake in **Question 2**.

To explore **Question 3**, I needed scenes where the target can be distinguished only by its type (Schemata A and B), scenes where a combination of type with

either colour or size suffices to describe the target (Schemata C and E), and scenes where all three non-relational properties are necessary (Schema D).

Note that the target can be described by its type, colour and size in all schemata; in other words, neither landmark nor distractor look identical to the target in any schema. This constraint ensures that the target's location in the scene or spatial relations to the other objects are never necessary to identify it.

Deriving Scenes from Schemata

To reduce the number of factors in the scene design, I varied the spatial arrangement of the three objects in the scene in only a few ways: The landmark and distractor are always placed clearly side by side, and the target is located either on top of or directly in front of the landmark. This results in four possible spatial configurations:

1. target on top of landmark, distractor to the left;
2. target on top of landmark, distractor to the right;
3. target in front of landmark, distractor to the left; and
4. target in front of landmark, distractor to the right.

The distractor is always placed slightly further away from the other two, with the aim of encouraging subjects to describe the target using its relation to the landmark, rather than its relation to the distractor. Whether the distractor is to the left or to the right of the other two objects is determined by the **orientation** of the scene. I did not expect the orientation of the scene to have an impact on the use of relations; however, the orientation is switched in half of the scenes each participant saw, in order to reduce monotony. Introducing the factor spatial relation by alternating the two different spatial relations between target and landmark also makes the scenes less monotonous and allows testing for the commonly observed cognitive preference of the vertical axis over horizontal axes (c.f., Lyons, 1977; Bryant et al., 1992; Gapp, 1995; Bryant et al., 2000; Landau, 2003; Arts, 2004; Tenbrink, 2004).

Each schema uses two colours, with the result that, in any scene, at least two objects have the same colour. Each scene uses one of two **colour templates**: blue+green or red+yellow. While I could of course not guarantee the exact hue, brightness and saturation displayed in a participant's browser, I used the four most prototypical colours in English to avoid naming issues. The two colour templates were distributed across the scenes such that this factor was balanced evenly with the factors **spatial relation**, **orientation**, and **scene schema**. This resulted in

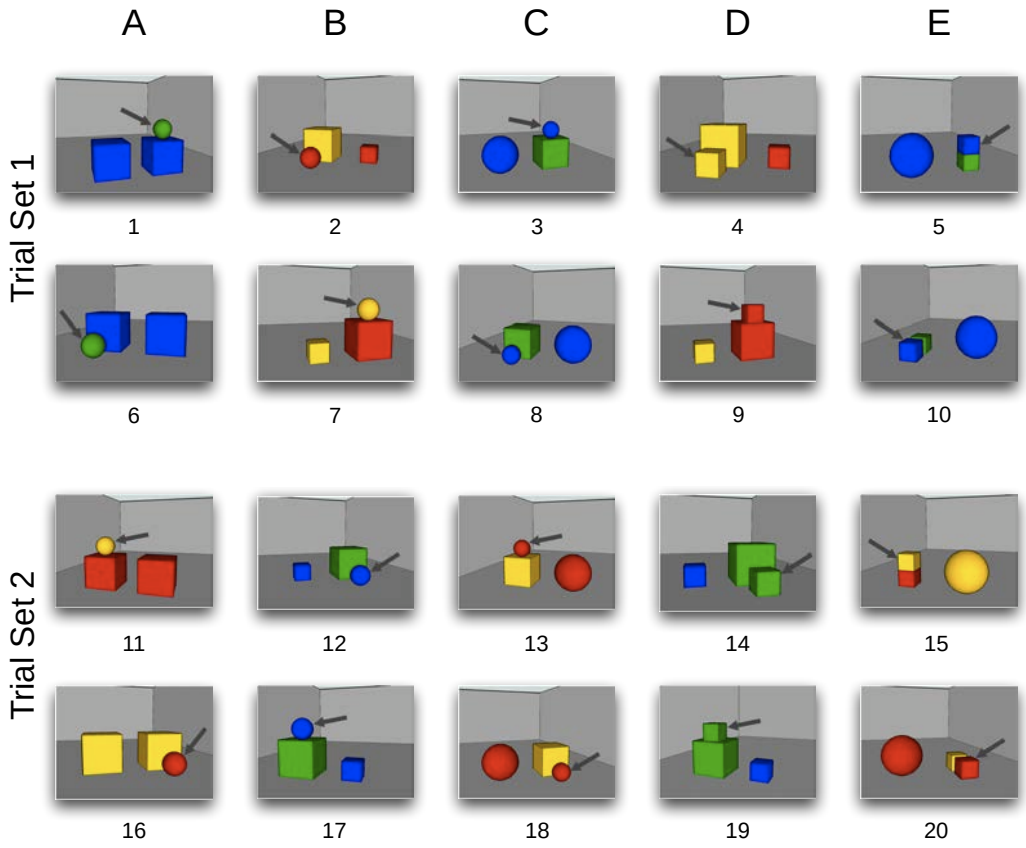

Figure 5.2: The GRE3D3 stimulus scenes. The letters indicate which schema from Figure 5.1 each column of scenes is based on.

six scenes being of one colour template and four of the other for each trial set. I did not expect the individual colours to influence which attributes were included in referring expressions; rather, I expect that this is influenced by whether the colours of each object are different from the colours of the other objects. To ensure that any unintended effects of the colour template apply to all conditions alike, the number of scenes using each colour template was balanced for each schema and for the type of spatial relation between the target and the landmark.

The 20 stimulus scenes are shown in Figure 5.2. I first created Scenes 1–5, each based on a different one of the five schemata. They were generated by alternating the colour template, the spatial relation between target and landmark and the orientation of the scene (i.e. the position of the distractor). I then created Scenes 6–10 by changing the spatial relation and orientation in each of the first five scenes. For example, Scene 8 was generated from Scene 3 by placing the target in front instead of on top of the landmark and flipping the scene vertically, so that the

distractor is on the right instead of the left. Scenes 1–10 constitute Trial Set 1. The second trial set, containing Scenes 11–20, was generated from the first one by changing the colour template in each scene and again flipping it along the vertical middle axis.

Because all 20 stimuli were generated from the same five schemata, they naturally fall into five different conditions. Due to the systematic generation process and the design principles outlined above, I ensured that the target–landmark relation, the orientation and the colour template of the scenes within each condition never fully coincide: if two scenes share one characteristic (e.g. the colour template), then they differ on the other two (orientation and target–landmark relation).

5.2.2 Procedure and Participants

The data gathering experiment for GRE3D3 was designed as a self-paced on-line language production study. Participants visited a website, where they first saw an introductory page with a set of simple instructions and a sample stimulus scene. Each participant was assigned one of the two trial sets containing ten stimulus scenes each. After the instruction page, the scenes were presented consecutively in a preset order. Below each scene, the participants had to complete the sentence *Please pick up the ...* in a text box before clicking a button labelled 'DONE' to move on to the next scene, as shown in Figure 5.3. The task was to describe the target referent in the scene (marked by a grey arrow) in a way that would enable a friend looking at the same scene to pick it out from the other objects.

To encourage the use of fully distinguishing referring expressions, participants were told that they had only one chance at describing the object. After being presented with all ten scenes in the trial, participants were asked to complete an exit questionnaire, which also gave them the option of having their data discarded and asked for their opinion on whether the task became easier over time and any other comments they might wish to make. A complete set of screenshots of the experiment webpages is provided in Appendix A.

74 participants completed the experiment. They were recruited by emailing self-reported native English speakers directly and asking them to pass on the invitation for participation. The participants were from a variety of different backgrounds and ages, but were mostly university-educated and in their early or mid twenties. For reasons outlined in Section 5.2.3 below, the data of 11 participants was discarded. Of the remaining 63 participants, 29 were female, while 34 were male.

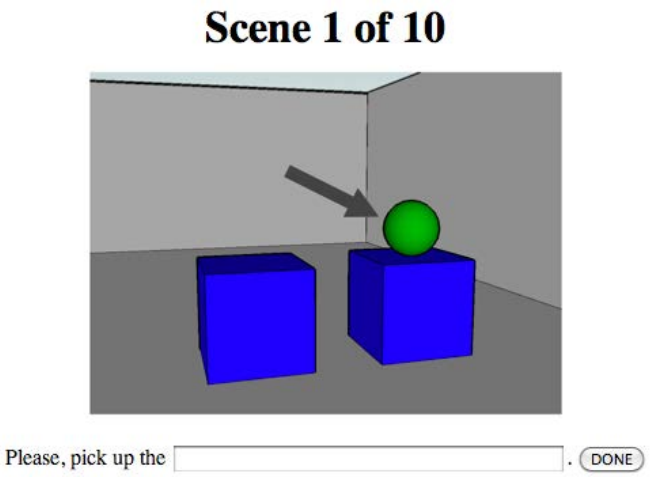

Figure 5.3: The screen showing the first GRE3D3 stimulus scene.

5.2.3 Data Filtering and Annotation

One participant asked for their data to be discarded, and I also disregarded the data of one other participant who reported to be colour-blind. One participant consistently produced very long and syntactically complex referring expressions including reference to parts of objects and the onlooker, such as *the red cube which rests on the ground and is between you and the yellow cube of equal size.* While these descriptions are very interesting, they were clearly extreme outliers in this data set and were therefore excluded from the analysis as well.

Eight participants consistently only used `type` to describe the target object, for example simply writing *cube* for the target in Scene 5. These descriptions were excluded from the corpus under the assumption that the participants had not understood the instructions correctly or were not willing to spend the time required to type fully distinguishing referring expressions for each trial. `type` alone was only distinguishing in 40% of the trials, those based on **schemata A** and **B**. While under-specified descriptions are justified in many real life situations, the task here is too straightforward to reasonably consider the use of under-specified descriptions sufficient. It is possible that this problem could have been avoided by pointing out more clearly to the participants that the (imaginary) person for whom they were producing the referring expressions could not see the marker pointing at the target object.

After removal of these data, 630 descriptions remain: 30 for each of the ten scenes from Trial Set 1, and 33 for each scene in Trial Set 2. The number of instances for each schema is 126.

In order to be able to analyse the semantic content of the referring expressions, I annotated the inherent attributes and relations contained in each of them. The attributes annotated are

- type [ball, cube]
- colour [blue, green, red, yellow]
- size [large, small]
- location [right, left, front, top][6]
- relation [on-top-of, in-front-of]

Each attribute is prefixed by either tg_, lm_, or dr_ to mark which of the objects it pertains to. For example, tg_size indicates that the size of the target was mentioned.

For this annotation, a limited amount of semantic normalisation was carried out: in the five cases where participants had used a more general relation such as *next to* or *adjacent to*, I annotated these with the specific relations that hold between the objects mentioned. Arguably, the difference between *next to* and *in front of* is of a semantic nature and not merely lexical; however, my analysis concentrates on whether the relation between two objects was mentioned, rather than how specific the value chosen for it was. For the same reason, I ignored the dynamic spatial preposition *from* in four descriptions such as the one in Example (5.1).

(5.1) the green ball from on top of the blue cube

The use of the dynamic preposition was most likely due to the movement implied by the indicated picking-up action, but, as already mentioned, I was mainly interested in the fact that the relation got used, not whether it was realised as a static or a dynamic preposition.

In descriptions containing comparatives, such as Example (5.2), I ignored the second object that the target was being compared to. In the context of the simple scenes at stake here, Example (5.2) is semantically equivalent to (5.3).

(5.2) the smaller of the two green cubes

(5.3) the small green cube

[6]Note that in Viethen and Dale (2008) we analysed locative expressions, such as *on the lefthand side*, as relations to regions of the scene and therefore counted them as spatial relations. Here, I treat them as a distinct class of properties, as they only indicate the general location of an object, rather than a spatial relation to another object.

	A	B	C	D	E	total
# descriptions	126	126	126	126	126	630
tg_col	85	81	78	124	123	491
tg_size	9	7	109	120	21	226
tg_loc	1	1	1	5	7	15
relation	47	46	59	38	34	224
lm_col	23	33	36	32	33	157
lm_size	2	18	4	38	7	69
lm_loc	35	3	3	3	2	46

Table 5.1: Attribute counts in GRE3D3. The number of descriptions that contain each attribute for each scene schema and in total.

Furthermore, how to deal with the relative nature of size is a separate, non-trivial, issue which is largely precluded from the studies in this thesis (but see, for example, van Deemter, 2000, 2006).

5.3 Analysis of GRE3D3

This section provides an analysis of the semantic content of the referring expressions contained in GRE3D3. In Section 5.3.1, I give a brief overview of the use of the different attributes contained in the corpus. Section 5.3.2 contains an in-depth analysis of the use of spatial relations that attempts to answer the research questions from Section 5.2.1.

5.3.1 General Overview

As discussed in Section 3.2.2, the type of the target referent is often included in a description even if it is not necessary for distinguishing the target from the other objects. In the GR3D3 Corpus, the type of the target is, in fact, included in all 630 descriptions. The type of the other two objects was included in all but two of the descriptions that mentioned these objects. The two exceptions are cases where the landmark was of the same type as the target object, which made it possible to replace the landmark's type by a *one*-anaphor, as in Example 5.4.

(5.4) the small blue cube which is on top of the green *one* [Scene 5]

Table 5.1 shows how often all other attributes were included in a referring expression in GRE3D3, both by scene schema and in total.

Despite the fact that spatial information was not necessary in any of the stimulus scenes to identify the target, 224 of the 630 descriptions (35.6%) contain a

spatial relation to the landmark. An example of a relational description that was given for the target in Scene 1 is given in Description 5.5.

(5.5) the green ball on top of the blue cube [Scene 1]

Ten of these relational descriptions additionally contain a spatial relation to the distractor object. In one of these ten descriptions the relation to the distractor describes the landmark (Description 5.6); in the nine other cases it relates the target object directly to the distractor (see Description 5.7 for an example).

(5.6) the smaller green cube located on top of the larger green cube which is to the left of a smaller blue cube [Scene 19]

(5.7) the small blue ball which is lying on top of the green cube and next to the big blue ball [Scene 3]

Ten instances do not constitute enough data to draw conclusions about the mention of a third object in a relational description. Therefore, I did not annotate the relation to the distractor in these ten descriptions, and they were not further analysed.

The target's colour (tg_col) was included in 491 (77.9%) of all descriptions, and 157 (70.1%) of the relational descriptions contain the landmark's colour (lm_col). 141 (67.4%) of the relational descriptions contain both tg_col and lm_col. The target's size (tg_size) was included in 266 (42.2%) of all descriptions, much less often than colour. The same trend is evident for the landmark's size (lm_size), which was only mentioned in 69 (30.8%) of the relational descriptions. 47 (23.7%) of the relational descriptions contain size for both objects.

Note that the scene schemata were designed in a balanced way such that tg_colour and tg_size were part of the inherent MD for the same number of stimulus scenes. In scenes based on Schemata A and B, neither tg_colour nor tg_size are necessary to describe the target referent, because in both the type of the target is unique. In scenes based on Schema C the inherent MD contains tg_type and tg_size, while for scenes based on Schema E tg_type and tg_col are minimally necessary to describe the target. The inherent MD in scenes based on Schema D contains both tg_col and tg_size in addition to the target's type. The difference in frequency of use can therefore not be attributed to an unbalanced need for the two attributes across the scenes in the stimulus set.

These counts are consistent with findings in the literature that indicate that colour gets included redundantly in referring expressions much more often than relative attributes such as size (see, for example, Belke and Meyer, 2002; Arts, 2004; Brown-Schmidt and Tanenhaus, 2006; Gatt, 2007). In fact, analysing only

the schemata in which each attribute was not necessary makes this picture much more pronounced: tg_col was used in 64.6% of cases in which it was not part of the inherent MD (scenes based on Schemata A, B and C), while tg_size was used in only 9.8% of cases in which it was not necessary to describe the target without using spatial information (scenes based on Schemata A, B and E).

Only 58 descriptions contain a locative expression that describes the location of an object within the scene. 15 descriptions mention the target's location (tg_location), while 46 descriptions mention the landmark's location (lm_location). Three descriptions use both the target's and the landmark's location, and only seven of the mentions of tg_location occur in descriptions without the spatial relation between the target and the landmark. The majority of mentions of lm_location occur for scenes based on Schema A, where the landmark looked identical to the distractor. It is interesting that many participants included location to distinguish the landmark from the distractor, despite the fact that the relation to the target already singled the landmark out uniquely. This appears to be evidence against the procedure employed by some REG approaches to spatial relations, which constrain the set of distractors of the landmark to those objects that stand in the same relation to the target as the landmark (cf. Krahmer and Theune, 2002; Siddharthan and Copestake, 2004).

5.3.2 The Use of Spatial Relations in GRE3D3

I now turn to the analysis of the use of spatial relations between the target and the landmark in the GRE3D3 Corpus. The main points of the detailed analysis in this section are summarised in Section 5.3.3. I first examine a number of general factors that might impact on the use of relations, including the colour template and orientation of the stimulus scene (which are not expected to have an influence), the type of relation holding between the two objects, and possible temporal effects. Following this, I analyse the impact of the target's and the landmark's visual salience on the use of the spatial relation between these two objects, in order to attempt to answer the research questions and hypotheses that formed the basis of the design of the stimulus scenes (see, Section 5.2.1).

General Factors

In Chapter 4, we saw that much of the variation in the Drawer Corpus could not be explained by the differences between the stimuli, as in some cases the same stimulus was described in many different ways. One possible explanation for such within-stimulus variation might be personal preferences of the different

participants who contributed to the corpus. Similarly, personal preferences might be the cause of at least some of the variation in the use of spatial relations in the GRE3D3 Corpus. The two most simple strategies a speaker might employ are to either always use a relation or never to use a relation, regardless of the referential scenario. The behaviour with regards to spatial relations of participants following one of these two exclusive strategies can be explained straightforwardly and only the data from the remaining participants would require further analysis.

In the GRE3D3 Corpus, more than half of the 63 participants used one of these two exclusive strategies: 11 participants opted to always use a relation, and 24 adopted a relation-free strategy, leaving 28 participants who varied their use of relations across the scenes. In the following, I concentrate on the analysis of the data from these 28 participants only. On average, these 28 non-exclusive strategists used a relation in 40.7% of their descriptions.

As expected, the colour template used in a scene did not have a significant effect on whether relations between objects were used. The difference between the two colour templates is less than five percentage points (χ^2=0.78, df=1, p>0.37)

As described in Section 5.2.1, half of the scenes displayed the distractor object to the left of the target–landmark cluster, while the other half had the distractor in the right of the scene. Just as the colour template, the orientation of the scene had no significant effect on the use of the spatial relation between the target and the landmark in the referring expressions produced (χ^2=0.06, df=1, p>0.81)

Figure 5.4 shows a very clear temporal effect in the use of the target–landmark relation in the GRE3D3 Corpus: a falling trend is evident from the first scenes the participants saw to the later ones. A χ^2 test between the data for the first five scenes and the data for the last five scenes shows that participants were significantly more likely to use a relation in the first half of the experiment than in the second half (χ^2=23.67, df=1, $p \ll 0.01$). On the basis of participants' comments provided on completion of the experiment, I believe that this decrease in the use of relations over time is due to a 'laziness effect', whereby subjects noticed after a few trials that relations were unnecessary and stopped using them. This suggests that the average use of relations would potentially be much higher in a setting where no such laziness effect could occur, such as, for example, in the first mention of an object in a real-world situation or in an experiment with filler items that do require relations for the identification of the target.

Recall that the target was placed either on top of the landmark or in front of it, with half of the GRE3D3 scenes (the odd-numbered ones) being of the first

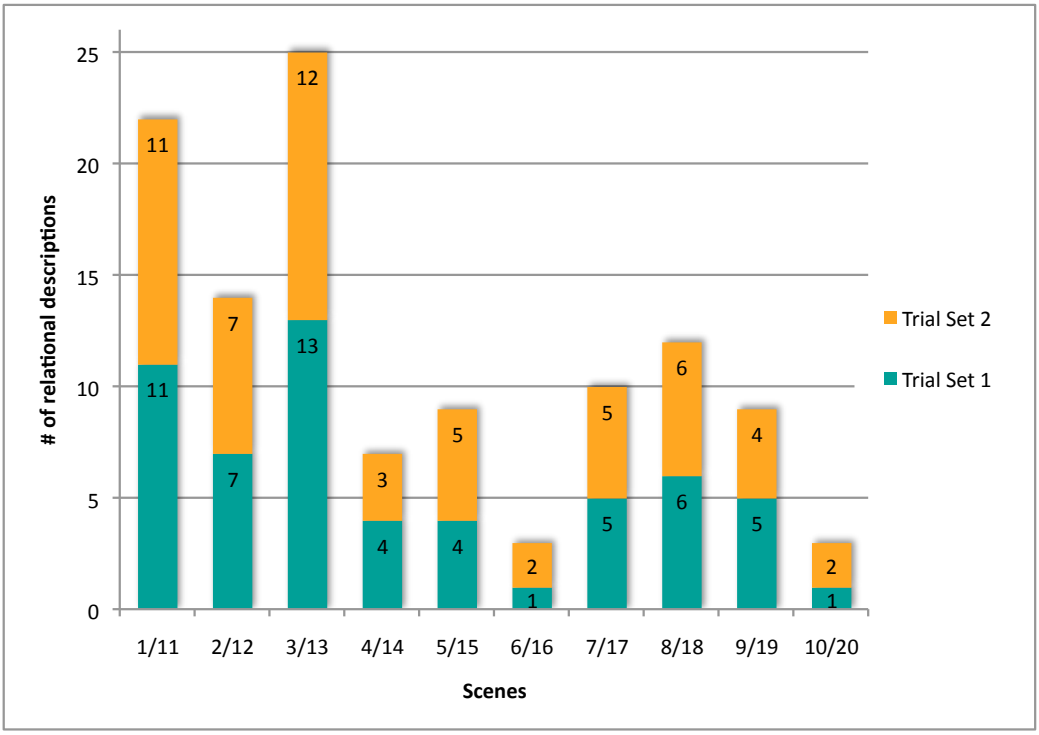

Figure 5.4: Number of relational descriptions by stimulus scenes. The counts for scenes from the different trial sets that only differ in colour template and scene orientation are stacked.

type, and the other half (the even-numbered ones) being of the second type. The psycholinguistic literature indicates that people prefer vertical relations over horizontal ones (for an overview, see Tenbrink, 2005, p.18), which is borne out by the data from the GRE3D3 Corpus: Of the 140 trials where the scene contained an on-top-of relation between target and landmark and which were described by participants not following an exclusive strategy, the on-top-of relation was mentioned in 75 instances (53.6%), while in-front-of was only used in 39 (27.9%) of the 140 'in-front-of' trials seen by non-exclusive strategists (χ^2=19.18, df=1, $p \ll 0.01$). This means that on-top-of relations were almost twice as likely to be mentioned as in-front-of relations by people who vary their use of relations between scenes.

Figure 5.5 shows that the on-top-of relation was used more often than in-front-of for every schema except Schema B. This exception might be due to the 'laziness effect' not having kicked in yet: The two in-front-of scenes for Schema B were Scenes 2 and 12, which got a relatively high number of relational descriptions (seven each), possibly because they were always displayed as the second scene in their respective trial sets.

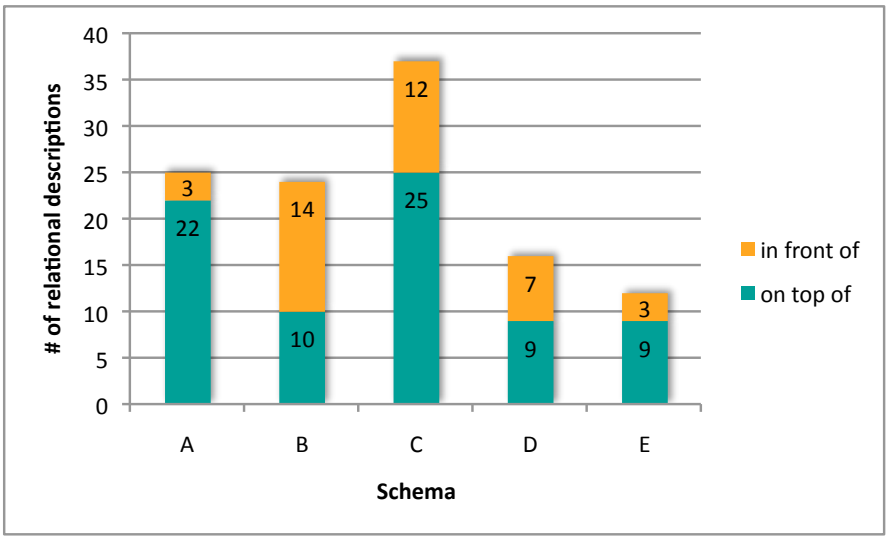

Figure 5.5: Number of relational descriptions for each schema.

The Impact of Object Similarity on the Use of Relations

In order to test **Hypothesis 1** from page 123, I have to find answers to **Questions 1** and **2** from Section 5.2.1.

Question 1 from Section 5.2.1 asked whether the use of relations is impacted by the similarity of the landmark to the target. **Hypothesis 1** expects that a landmark sharing few properties with the target would result in relations being used more often due to the landmark's higher visual salience. The impact of landmark–target similarity can be tested by comparing the descriptions given for scenes based on Schemata D and E, where target and landmark share two properties, to those for scenes based on Schemata A, B and C, where target and landmark share no properties.

A clear effect was found between these two conditions: The participants who did not follow an exclusive strategy were twice as likely to use a relation for scenes with dissimilar target and landmark (49.4%) than for the scenes where they were similar (24.1%, χ^2=86.29, df=1, $p \ll 0.01$). This outcome suggests that, at least in the GRE3D3 domain, a visually salient landmark increases the likelihood of the relation to this landmark being included. However, this result has to be viewed with some caution, because the scenes based on Schemata D and E are the ones that are likely to be affected most by the laziness effect reducing the use of relations.

As a second component of the landmark's salience in the GRE3D3 scenes I investigate its similarity to the third object, the distractor. The landmark–distractor

similarity was at stake in **Question 2** from Section 5.2.1: does the similarity between landmark and distractor have an impact on the use of the relation between the target and the landmark? Again, **Hypothesis 1** expects that the more similar the landmark is to the distractor, the less salient it is, and the less likely it is therefore to be used in a referring expression.

There was almost no difference between the usage of relations for scenes where landmark and distractor are identical (based on Schema A) and those where they share one property value (based on Schemata B, C and D). When they were identical, relations were used in 44.6% of descriptions, and when they shared one property value in 45.8% of descriptions.

However, the use of spatial relations dropped significantly for scenes where the distractor is completely distinct from the landmark object (Schema E). 21.4% of all descriptions for these scenes contained relations (χ^2=10.79, df=1, $p \ll 0.01$). The hypothesis underlying Question 2 was therefore not confirmed. Schema A, where the landmark is identical to the distractor, did not result in a lower rate of relation use, and Schema E, where the landmark had nothing in common with the distractor, received a lower, rather than the expected higher, rate of relation use than the other schemata. Again, it is possible that the low use of relations for scenes based on Schema E is mostly due to the influence of the laziness effect, as these scenes were always the fifth and tenth scene a participant saw. Similarly, the average use of relations for Schemata B, C and D is lowered by the low results for Schema D, which also might be affected by the laziness effect. This might explain why the rate of relation use for Schema A is relatively high in comparison.

Another way of testing for an effect of the landmark's salience and **Hypothesis 1** is to consider the landmark's similarity to both other objects at the same time. In scenes based on Schema C, the landmark is unique from the other two objects both in type and in colour, while in all other scenes it shares at least its type with another object. Figure 5.6 shows that the average use of relations is much higher for scenes based on Schema C than for scenes in which the landmark is either similar to the target (those based on Schemata D and E) or to the distractor (scenes based on Schemata A and B). In 66.1% of all trials where the scene contained a landmark very dissimilar to both other objects (Schema C), a spatial relation to the landmark was used by the participants not following an exclusive strategy, while only 34.4% of the descriptions for the other schemata contained a relation. This difference is highly statistically significant (χ^2=18.65, df=1, $p \ll 0.01$).

Considering that the target object can be assumed to be the main focus of at-

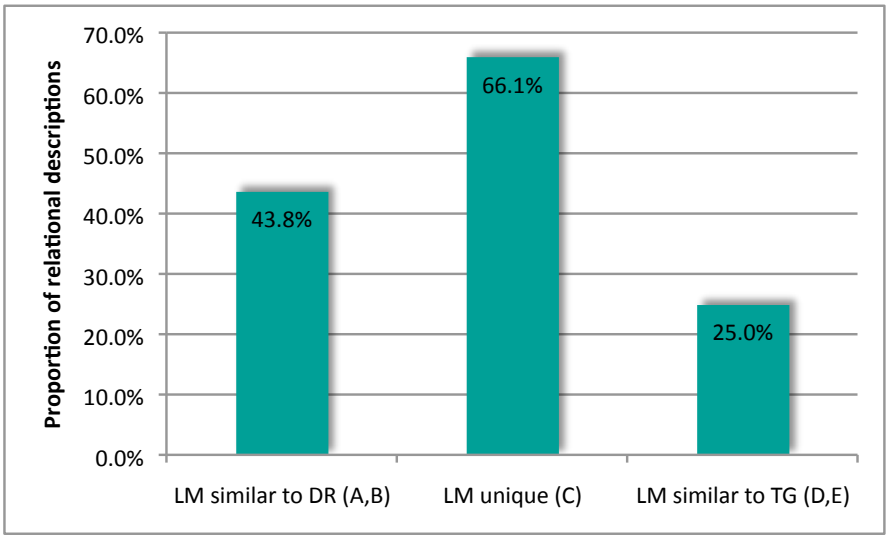

Figure 5.6: Impact of the similarity of the landmark (LM) to target (TG) and distractor (DR) on the use of relations.

tention of a person producing a referring expression, the visual differences between the landmark and the target might be of more importance than those between the landmark and the distractor. The target–landmark relation was used in 43.8% of the scenes where the landmark was more similar to the distractor than the target (Schemata A and B). On the other hand, landmarks sharing more properties with the target than with the distractor (Schemata D and E) were included in referring expressions in only 25.0%. The difference between these two conditions is statistically significant at $p<0.01$ ($\chi^2=8.08$, $df=1$). This lends support to the results from above which showed that a landmark similar to the target is less likely to be included in a referring expression than one that is visually different from the target, while a higher similarity between the landmark and the distractor makes the landmark more likely to be included.

In summary, **Hypothesis 1** has been confirmed: relations to landmarks that are visually very different from the other objects in the scene, and can therefore be assumed to be more visually salient, are more likely to be used in a referring expression than landmarks that are similar to one of the other objects.

Hypothesis 2 from page 123 states that a target with a short inherent MD should be less likely to be described in terms of a spatial relation than a target whose inherent MD is longer. In order to test this hypothesis and thereby answer **Question 3** from Section 5.2.1, I tested for the influence of the number of

properties shared by the target with the other objects in the scene, which can be considered a rough measure of the target's visual salience.

In Schemata A and B, the target does not share its type with either of the other objects, which means it can be described uniquely by using only type, resulting in an inherent MD of length one. In Schemata C and E, it shares its type and one other properties with another object, so the length of its inherent MD is two: type and the property not shared with the object of the same type. While not being identical with either object in Schema D, here the target shares each of its three properties with at least one of the other objects and therefore can only be distinguished by a referring expression of at least length three.

The proportion of descriptions that used a relation to the landmark was 43.8% for schemata where the inherent MD is of length 1 and length 2, and 28.6% for targets whose inherent MD is of length 3. However, this difference is not statistically significant (χ^2=4.28, df=2, p>0.1) and **Hypothesis 2** can therefore not be confirmed.

It seems that these results are heavily influenced by a different factor which is masking any effect of this factor: To test for the effect of the length of the target's inherent MD I had to pool the results for Schema C with those for Schema E, which are the schemata with the highest and lowest use of relations overall. It is likely that any effect of a factor based on the similarity between these two schemata is overshadowed by a much stronger effect of another factor based on the difference between them.

5.3.3 Interim Summary

The GRE3D3 Corpus was collected in order to gain some initial insights into the way people use spatial relations in referring expressions. The stimulus scenes were designed in a way that allows tests that might answer three research questions. These questions were based on two hypotheses: (1) that the spatial relation between the target object and a landmark object is more likely to be included in a referring expression if the landmark object is visually salient in the scene, and (2) that the target–landmark relation is less likely to be included if the target object can be described easily by a short non-relational referring expression (inherent MD). The first of these hypotheses was confirmed, as the relation to the landmark object was most often included for scenes in which the landmark was maximally different in its visual properties from the other objects in the scene. The second of these hypotheses was not confirmed by the GRE3D3 data.

The following observations can be taken away from the analysis of the GRE3D3 Corpus:

1. *People use spatial relations even if the target referent can be distinguished from all distractors by its inherent visual properties alone.* More than a third of the descriptions in the GRE3D3 Corpus contained a spatial relation despite the fact that the targets in all stimulus scenes could be described without relations. Most REG algorithms that can deal with relations between entities only include them if absolutely necessary. These algorithms are, therefore, ruled out as potential models of human reference behaviour.

2. *Much of the variation in the use of spatial relations is due to participant-specific preferences.* About 55% of the participants in the GRE3D3 collection experiment adopted one of two exclusive strategies by either including a relation in all their referring expressions or never using a relation. Only 28 participants varied their reference behaviour with respect to relations between the different scenes they saw. This suggests that REG algorithms aiming to replicate human data or even explain human reference behaviour need to pay more attention to participant-specific effects than existing approaches do.

3. *People are more likely to use a spatial relation in a one-off reference situation.* Once they get to know a domain and find that relations are not necessary, they become 'lazy' and opt for shorter, relation-free descriptions. The data in the GRE3D3 Corpus shows a clear downward trend in the use of relations towards the end of the experiment. This suggests that if each instance had been a one-off description, rather than part of a chain of ten back-to-back referential scenarios, the use of relations would have been even higher. Unfortunately, this laziness effect obscures to some extent the results of analyses of other influencing factors.

4. *Relations in the vertical axis are preferred over those in a horizontal axis.* In accordance with previous findings in the psycholinguistic literature, I found that people prefer to use relations in the vertical axis over horizontal relations. Target referents placed on top of the landmark object were almost twice as likely to be described in terms of the relation to this landmark than targets sitting in front of the landmark object.

5. *The more visually salient a landmark is, the more likely it is to be included in a referring expression via its spatial relation to the target.* Visual salience is determined here by the degree to which an object differs in its visual properties from the other objects in the scene. Relations to landmarks that

were very different from the other two objects in the scene were much more likely to be included in a referring expression.

As I mentioned in the introduction to this chapter, the GRE3D3 Corpus was intended as a pilot study for a more comprehensive experiment. In the second data collection experiment, presented in the second half of this chapter, I attempted to apply the lessons learned from the GRE3D3 experiment. The second experiment was designed to minimise the laziness effect I observed in the GRE3D3 data and to elicit in more detail how the salience of the landmark object might influence the use of relations. It is focussed, in particular, on the impact that the landmark's size has on the use of spatial relations.

5.4 Collecting GRE3D7

5.4.1 Stimulus Design

The stimulus scenes used for the GRE3D7 corpus were similar in many regards to those used for GRE3D3. They were three-dimensional scenes created in Google SketchUp containing only simple geometric shapes. In GRE3D7, each stimulus scene contained seven objects, making the scenes slightly more complex than the GRE3D3 scenes. The seven objects were grouped into three pairs of two and one single object. The target object was always part of one of the pairs and the second object of that pair is what I call the **landmark** object in these scenes. I attempted to place the target–landmark pair as close to the centre of the scene as possible to discourage the use of overall location in the scene such as 'in the left'. The other two object pairs were placed slightly further back to the left and right of the target–landmark pair and the single object was always placed in the far right or the far left of the scene. As in GRE3D3, objects were either balls or cubes and otherwise distinguishable by their size and colour. Each object could be either large or small, and in each scene I use only two colours.

The two main hypotheses underlying this data collection exercise are concerned with the influence of the landmark object's size on its salience and the likelihood of the target–landmark relation being used in a referring expression:

> **Hypothesis 3:** A large landmark is more salient than a small one because it occupies more of the visual space of a scene. Therefore, a large landmark is more likely to be mentioned in a referring expression via its spatial relation to the target referent than a small landmark.

Hypothesis 4: A landmark that shares its size with a number of other objects in the scene is less salient than one that is unique in size. Therefore, a landmark with unique size is more likely to be mentioned in a referring expression via its spatial relation to the target referent than a landmark with a common size.

A second consideration that might influence the use of relations, apart from the landmark's overall salience in the scene, is the similarity between the target and the landmark object. At the time when the landmark's salience is taken into account, the participants are focusing their attention on the target object. As the landmark is the closest object to the target, it is likely that the difference or similarity between these two objects plays a particularly important role in the decision whether to include the relation between them or not. Two conflicting hypotheses can be formulated here:

Hypothesis 5: The difference between the landmark and the target object impacts on the visual salience of the landmark because it impacts on the landmark's overall uniqueness in the scene. Therefore, a landmark that is visually different from the target is more likely to be included in a referring expression than one that looks similar to the target.

Hypothesis 6: The more similar the landmark and target objects are, the more they appear as one visual unit rather than two separate objects. If they are perceived and conceptualised as a visual unit, they are more likely to be mentioned together. Therefore, the more similar the landmark is to the target, the more likely it is to be included in a referring expression.

The fifth hypothesis that this experiment is designed to test concerns the preference that participants in the GRE3D3 Corpus, as well as previous psycholinguistic work, showed for vertical relations over horizontal ones. To make sure that the landmark is never obscured by the target object, I use lateral relations rather than frontal ones in this experiment.

Hypothesis 7: A target placed on top of a landmark object is more likely to be described in terms of its spatial relation to the landmark than a target that is sitting directly adjacent to the left or right of the landmark.

I report the results for the tests of these five hypotheses to the test on pages 151 to 152 in Section 5.5.2. To be able perform these tests systematically, the experiment was designed as a 2×2×2×2×2 grid with the following five variables:

- LM_Size: the landmark is either large or small. [Large/Small]
- LM_Size_Rare: the size of the landmark is either a common size in the scene or it is as rare as possible, even unique. If it is common and the landmark is large it shares its size with two of the objects, if it is small with three. These numbers are not the same because in each scene in which the landmark size was common three objects were large and four small. In +LM_Size_Rare scenes that are also +TG_Size=LM_Size the landmark shares size only with the target. Only if the scene is –TG_Size=LM_Size can the landmark's size be truly unique in the scene. [+/–]
- TG_Size=LM_Size: target and landmark are either the same size or different. [+/–]
- TG_Col=LM_Col: The target and the landmark were either of the same colour or of different colour. [+/–]
- Relation: The relation between the target and the landmark is either vertical (the target is on top of the landmark) or lateral, in which case the target is placed directly to the left or right of the landmark. [Vertical/Lateral]

This resulted in 32 experimental conditions. I created one stimulus scene for each of these conditions. I then split the stimuli into two trial sets along the factor TG_Size=LM_Size, so that this variable became a between-participant factor, while the other four are within-participant factors.

There were a number of other criteria that I followed for the design of the stimulus scenes to minimise the possibility of unwanted factors influencing the results:

Target uniqueness: The target was always unique with its inherent properties alone, which means that the relation to the landmark or other external properties, such as the location in the scene, were never necessary to fully distinguish the target from all other objects in the scene.

Landmark uniqueness: As the target, the landmark was always unique with its inherent properties alone.

Colour balance: As for GRE3D3, each scene followed one of two colour schemes: either blue–green or red–yellow. The colour schemes were distributed in a balanced way across the five experimental variables, so that half of the scenes in each condition were blue–green and the other half red–yellow. The colour scheme was not expected to have an influence on the content of the referring expressions people produced. In each scene, four objects were of one colour of the colour scheme for this scene and three had the other colour.

Relation balance: The relation between the target and the object was never unique. One of the two other object pairs in each scene had the same spatial relation as the target–landmark pair and the third pair had the other relation. However, the objects in the pair with the same relation were never of the same types as target and landmark, so that a description of the form ⟨tg_type, relation, tg_type⟩ was always fully distinguishing.

Constant landmark and target types: The landmark was always a cube to avoid scenes where the target would have to be balanced on top of a sphere, which might look unnatural. The target was always a ball to make sure that the similarity in type between these two objects was always constant.

No obscured objects: The objects were placed in the scenes in such a way that no object occluded any other. In particular, as mentioned above, there were no frontal relations within the object pairs, to avoid larger objects obscuring smaller ones completely or to a large degree.

Figure 5.7 shows the 2×2×2×2×2 grid of the 32 stimuli scenes. Scenes 1–16, shown on a green background, constitute Trial Set 1, and Scenes 17–32, shown on a blue background, constitute Trial Set 2.

5.4.2 Procedure and Participants

The data collection procedure was in most respects identical to that for the GRE3D3 Corpus; however, there were two important differences. Firstly, the order in which the stimulus scenes were presented was randomised for each participant; and secondly, before each of the 16 stimulus scenes, the participants were shown a filler scene, which means each participant had to describe 32 scenes in total. The main motivation for making these changes to the experimental design was to minimise the decline in relation use over time due to the laziness effect observed in the GRE3D3 collection experiment.

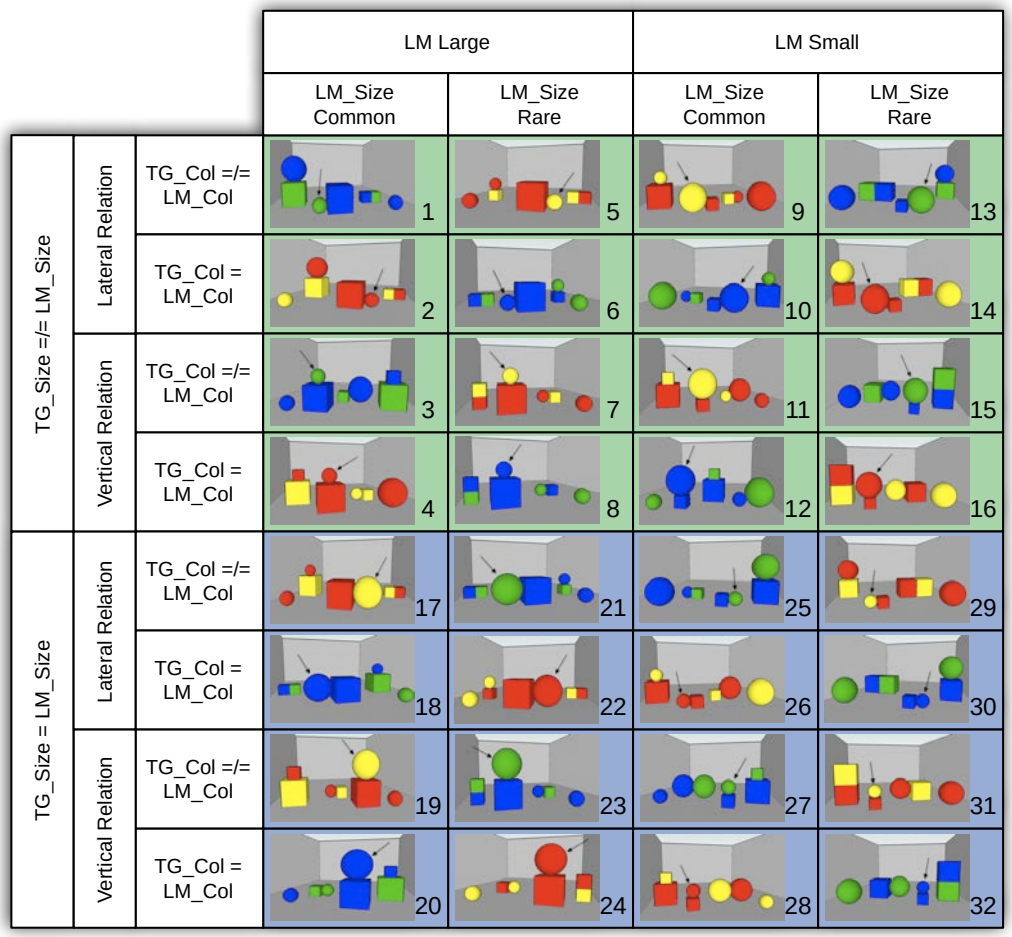

Figure 5.7: The 32 stimulus scenes for GRE3D7. The top half constitutes Trial Set 1 and the bottom half is Trial Set 2.

The filler scenes were also designed with the intention of making the experiment less boring and to stop participants from noticing the strict design features of the stimulus scenes. In particular, each participant saw four scenes with twelve objects in all four colours rather than adhering to a two-colour scheme, two scenes from Trial Set 1 from the GRE3D3 stimuli containing only three objects, and ten further filler scenes which intentionally violated the above design criteria. The filler scenes each participant saw were chosen such that in eleven or twelve scenes the target was a cube instead of a ball, in two scenes the landmark was a ball, in four scenes there was no obvious landmark close to the target, in eight scenes the target was unique (i.e. it could not be described by its inherent visual properties alone), in nine or ten scenes the target and landmark shared type, and in two or three scenes target and

landmark were of the same size (for participants who saw Trial Set 2 all stimulus scenes of course also had a target and landmark of the same size). Screenshots of the experiment webpages and all filler scenes are shown in Appendix B.

The sequence of the 32 scenes that were shown to a particular participant was determined by the following three steps:

1. pick the opposite trial set to the one that the last participant saw and randomise its order,
2. pick the set of 16 filler scenes to be shown to this participant and randomise their order,
3. interleave the two sets so that each stimulus scene is preceded by one filler scene.

318 people started the experiment, of which 294 participants completed all 32 scenes. They were recruited in the same way as the participants for GRE3D3, but the call for participation was circulated more widely and was also published on two electronic mailing lists[7]. The participants were predominantly in their twenties or thirties and mostly university-educated. A slight majority were female (54% vs. 46% male).

5.4.3 Data Filtering and Annotation

Of the 294 participants who completed the experiment, five consistently used only type, although the target's type was never fully distinguishing in any of the stimulus scenes. I discarded the data of these participants under the assumption that they had not understood the instruction that their descriptions were to uniquely identify the target. Two participant's data were discarded because they wrote things unrelated to the displayed scenes. Of the remaining 287 participants, 140 saw Trial Set 2 and 147 saw Trial Set 1. The data from seven participants from Trial Set 1 were discarded to balance the corpus in terms of the between-participant feature TG_Size=LM_Size. Each person described the 16 scenes contained in either of the trial sets, resulting in a corpus of 4480 descriptions in total, 140 for each scene.

Only five of the 4480 descriptions used the ternary spatial relation *between*, and one description mentioned two separate spatial relations, one to the intended landmark and one to another object. The relation to the third object in these six descriptions was disregarded in the analysis.

[7] `http://www.hit.uib.no/corpora` and `http://www.siggen.org`

attribute	count	% of total 4480 descriptions	% of all 600 relational descriptions
tg_size	2587	57.8	–
tg_colour	4423	98.7	–
tg_location	81	1.8	–
relation	600	13.4	–
lm_size	327	7.3	54.5
lm_colour	521	11.6	86.8
lm_location	10	0.2	1.7

Table 5.2: Attribute counts in GRE3D7

I annotated the data in the same way as described for the GRE3D3 Corpus in Section 5.2.3.

5.5 Analysis of GRE3D7

5.5.1 General Overview

The target object's type was mentioned in each description in the GRE3D7 Corpus and each relational description contained the landmark object's type. Table 5.2 shows how many descriptions contained each of the other attributes.

The Sparing Use of location

Only 81 (1.81%) descriptions made reference to the target referent's location in the scene, as in Example (5.8), and ten of the 600 relational descriptions (1.67%) contained the location of the landmark, as in Example (5.9). There were no descriptions containing both tg_location and lm_location. This might indicate that in the GRE3D7 Corpus participants who used a relation were more likely to conceptualise the target–landmark pair as a unit with just one location rather than as two individual entities. However, neither of the corpora were designed to investigate this issue and the numbers for use of location are too low in both corpora to draw any definite conclusions.

(5.8) the large yellow ball on the left [Scene 9]

(5.9) the ball small next to the large cube on the left hand side [Scene 6]

The Abundant Use of colour

Colour was used in the vast majority of descriptions. 98.7% of all descriptions included the colour of the target object and 86.8% of the relational descriptions included the colour of the landmark object. A high number of descriptions containing colour was expected, as colour was part of the inherent MD for 20 of the 32 scenes (all but Scenes 17–24 and 29–32). However, the fact that colour was also included in the majority of the descriptions containing spatial information, in the form of a relation or the location, confirms the trend in GRE3D3 (see Section 5.3.1) whereby colour is often included in descriptions redundantly, as well as previous findings to this effect (Belke and Meyer, 2002; Arts, 2004; Gatt, 2007).

The Utilitarian Use of size

The target's size was mentioned in 57.8% of all descriptions, and the landmark's size in 54.8% of the relational descriptions, both more often than in the GRE3D3 Corpus, where the respective percentages were 42.2% and 30.8%. The difference in use of tg_size in the two corpora is interesting, as the proportion of stimulus scenes in which tg_size was part of the inherent MD is very similar for both corpora: 12 of 32 scenes (37.5%) in GRE3D7 (Scenes 2, 4, 9–12, 18, 20 and 25–28) and 8 of 20 scenes (40%) in GRE3D3 (scenes based on Schemata C and D). The difference is statistically significant at $p \ll 0.01$ ($\chi^2 = 352.20, df = 1$)[8].

The use of tg_size for scenes where it was part of the inherent MD was almost identical in the two corpora (90.8% for GRE3D3 and 90.2% for GRE3D7). The difference stems therefore from the scenes where tg_size is not necessary in a description using only inherent visual properties; targets in the GRE3D7 scenes where tg_size was not part of the inherent MD were more likely to be described in terms of their size than in the equivalent GRE3D3 scenes.

Let us therefore consider in particular the scenes where tg_size was not part of the inherent MD. A possible explanation for the difference in the use of tg_size for scenes of this kind in the two corpora might lie in the differing usefulness of tg_size. The GRE3D3 scenes of this kind (scenes based on Schemata A, B and E) contained no object that shared type but not size with the target. In other words, once tg_type has been included in a referring expression for the target in one of these scenes, the discriminatory power of tg_size is 0. In GRE3D7, on the other hand, 12 of the 20 scenes where tg_size was not necessarily part of the inherent MD

[8] In this calculation of χ^2 the proportion of scenes in which tg_size is part of the inherent MD is used as the expected value for each corpus, rather than the mean of the actual occurrence of tg_size.

(Scenes 1, 3, 5–8, 13–16, 17, 19, 21–24 and 29–32) nonetheless contained another object that shared the target's type (ball) but not its size (Scenes 1, 3, 17, 19, 21–24 and 29–32). In these scenes, tg_size remains a useful attribute to use, even after tg_type is already included.

One might expect that the use of tg_size is higher for these GRE3D7 scenes because here it helps distinguish from another object of the same type rather than only from objects of a different type. This hypothesis is supported by the data: tg_size was used in 45.6% of the GRE3D7 descriptions for scenes where it was not part of the inherent MD but there was another object of same type and different size as the target. For scenes where tg_size could only distinguish the target from objects of the other type, it was only used in 27.3% of cases (χ^2=94.97, df=1, $p \ll 0.01$). This shows that size is mostly used to compare to or distinguish from other objects of the same type, while the same is not true for colour. This finding is in accordance with findings from eye-tracking experiments in psycholinguistics (c.f. Sedivy, 2003; Brown-Schmidt and Tanenhaus, 2006)

5.5.2 The Use of Spatial Relations in GRE3D7

600 of the 4480 descriptions in the GRE3D7 Corpus (13.4%) mentioned a spatial relation. This number is significantly lower than in the GRE3D3 Corpus, most likely because each scene contained another object pair with the same spatial relation as the one holding between the target–landmark pair. In contrast, the GRE3D3 scenes only contained one object pair, giving the spatial relation itself more discriminatory power and making it more visually salient. However, we have to keep in mind that spatial information was not required in any of the stimulus scenes. Most existing approaches to spatial relations in REG would therefore never include a relation for any of the stimuli.

In this section, I examine the circumstances under which the participants of the GRE3D7 data collection experiment used the spatial relation between the target object and the intended landmark. As I did for GRE3D3, I will first examine participant-dependent and temporal factors and then move on to analyse the impact that the design features of the scenes, described above in Section 5.4.1, had on the use of relations.

General Factors

I first checked for broad participant-dependent preferences for or against using relations in the GRE3D7 Corpus. The behaviour of participants who use an exclusive

strategy of either always or never including a relation in their referring expressions is easy to predict and does not contribute to any variation across different scenes. In order to gain a clear understanding of this variation, I will concentrate on the data from participants who varied their use of relations between scenes.

The proportion of participants who adopted an exclusive strategy regarding the use of relations was similar for the two corpora at 50.3% for GRE3D7 and 55.5% for GRE3D3. However, the split between the two exclusive strategies was much more uneven in GRE3D7: 135 participants never used a spatial relation and only six used a spatial relation for all 16 stimulus scenes they saw, compared to a 24–11 split in GRE3D3. Of course, it is likely that the high proportion of participants choosing to never use a relation is, to some extent, due to a difference between the stimulus scenes for the two corpora, rather than being only participant dependent. I will discuss these below. In the following, I analyse the data from the 139 participants who used a relation for some scenes but not for others. On average, these participants used a relation in 22.7% of their descriptions.

In GRE3D3, I observed a 'laziness effect' whereby participants' use of relations decreased over the course of the experiment. A number of participants mentioned in the exit interview that they noticed over time that relations were never required and stopped using them. Such a conscious, or semi-conscious, adjustment masks people's natural propensity to use a relation in a reference situation where they come anew at the task rather than describing one object after another.

In the GRE3D7 collection experiment, each participant saw eight filler scenes in which spatial relations were required to distinguish the target. These filler scenes were included to stop participants from consciously noticing that relations were never required in the stimulus scenes. I hoped that this would reduce the laziness effect and thereby better approximate people's natural tendency to use a relation. However, Figure 5.8 shows that, despite the use of these filler scenes, the use of relations declined over the course of the experiment. Participants who did not follow an exclusive strategy clearly used more relations for scenes they saw early on than for those they saw towards the end. I divided the data set into quartiles in order to test the statistical significance of this decline. Figure 5.9 shows the average proportion of descriptions in each quartile that contained a spatial relation. The falling trend was statistically significant at $p \ll 0.01$ ($\chi^2=55.42$, $df=3$).

However, the use of the filler scenes did have some of its intended effect: the decline in use of relation was less pronounced in GRE3D7 than it was in GRE3D3, where people were more than twice as likely to use a relation in the first half of the experiment compared to the second half. In GRE3D7, the use of relations

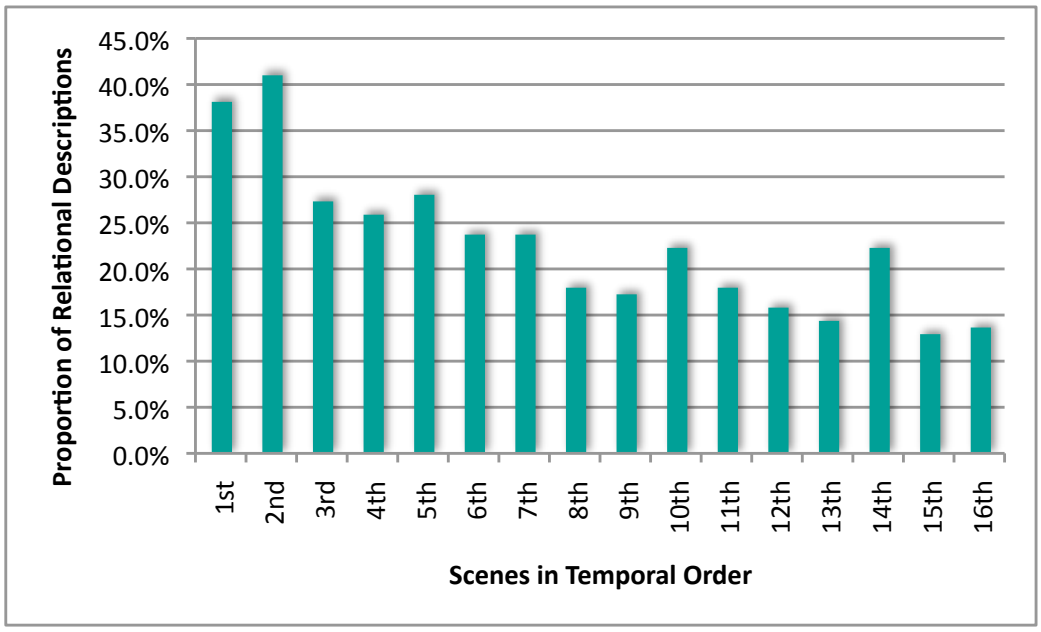

Figure 5.8: Temporal Effect on Use of Relations in GRE3D7.

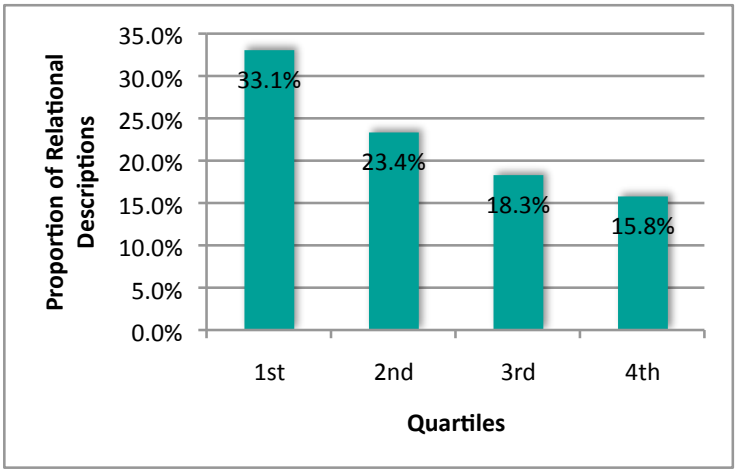

Figure 5.9: Use of Relations in GRE3D7 by Quartile.

was only 1.6 times higher in the first half than in the second. Furthermore, any temporal effect in GRE3D7 should not interfere with between-stimulus effects, as the stimuli were presented in a randomised order.

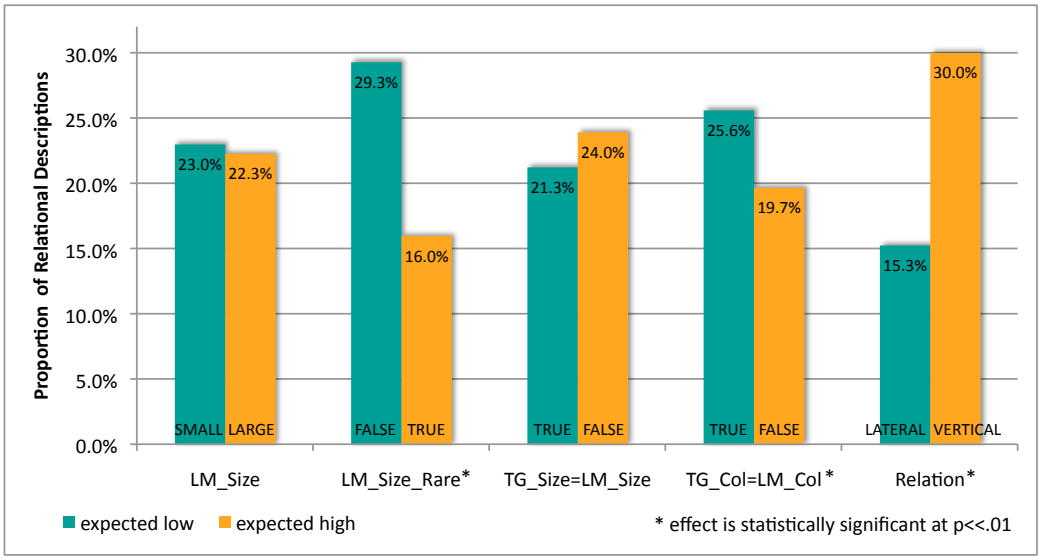

Figure 5.10: The Effect of the Design Variables on the Use of Relations in GRE3D7

Influence of Scene Features on the Use of Relations

I will now turn to the examination of **Hypotheses 3–7** from Section 5.4.1. Figure 5.10 shows the impact that each of the five variables of the scene design had on the use of relations. The left (green) columns represent the conditions for which I expected less relations to be used and the right (yellow) columns represent the conditions for which I expected a higher use of relations, according to **Hypotheses 3–5** and **7**. **Hypothesis 6** expected the reverse results for TG_Size=LM_Size and TG_Col=LM_Col. So, for example, I expected a higher use of relations for scenes with small landmarks than for those where the landmark was large; but the use of relations was at 23% and 22.3% almost the same for these two conditions. All factors except LM_Size and TG_Size=LM_Size had a statistically significant effect.

Hypotheses 3 and **4**, which expected a large landmark with a rare or unique size to be more salient and therefore more likely to get used, are not supported by the data here. LM_Size did not have a reliable effect and LM_Size_Rare shows the opposite effect of the one I expected: a relation to a landmark with a common size is significantly more likely to be included in a referring expression than one to a landmark with a rare or unique size. On closer inspection, it transpires that this is likely to be due to a factor that was not explicitly tested in this experiment: the length of the inherent MD of the target referent. In most scenes with a common

landmark size (all but Scenes 1, 3, 17, and 19), all three inherent attributes, size, colour and type, are necessary to distinguish the target from the other objects without using locational information. In all scenes where the landmark's size is rare or unique, colour and type suffice. In other words, targets which are harder to describe using inherent visual properties only are more likely to be described by a relation to a nearby landmark. This result confirms **Hypothesis 2** from Section 5.2.1, for which no conclusive answer could be found in the GRE3D3 Corpus.

Hypotheses 5 and **6** predicted two mutually exclusive scenarios based on the assumption that the similarity between the target and the landmark object is of special importance, as the participants' visual attention is likely to be focussed on these two objects. **Hypothesis 5** predicted that a visual difference between the landmark and the target would increase the landmark's salience and therefore the use of the spatial relation to this landmark. **Hypothesis 6** predicted that high visual similarity between target and landmark might result in these two objects being conceptualised as a unit which would increase the likelihood of both objects being mentioned. The target and landmark object were always of different types, so their similarity depends on their size and their colour, captured in the variables TG_Size=LM_Size and TG_Col=LM_Col. TG_Size=LM_Size did not show a significant effect on the use of relations ($p>.1$). The effect of TG_Col=LM_Col favours **Hypothesis 6**, as a landmark of the same colour as the target is more likely to be included in the target's description than one that has a different colour from the target.

The variable Relation had the expected effect: A vertical relation is significantly more likely to be used than a lateral one. This confirms **Hypotheses 7**.

Figure 5.11 visualises the use of relations for each individual scene. The darker the background behind a scene, the higher the proportion of relational descriptions that were produced for this scene. Scene 2 received at 56.9% the highest proportion of relational descriptions and Scene 14 at 2.8% the lowest. The pattern of shading indicates an interaction between the two factors that each by themselves had no reliable effect on the use of relations, LM_Size and TG_Size=LM_Size. If target and landmark have the same size, small landmarks are more likely to be used in a referring expression (χ^2=14.94, df=1, $p\ll$0.01); but if target and landmark are of different sizes, large landmarks get mentioned more often (χ^2=22.82, df=1, $p\ll$0.01). This translates into a higher use of relations in scenes with a small target (28.1%) than in those where the target is large (17.3%)(χ^2=63.94, df=1, $p\ll$0.01).

Looking at the effect of LM_Size in interaction with the target's size reveals

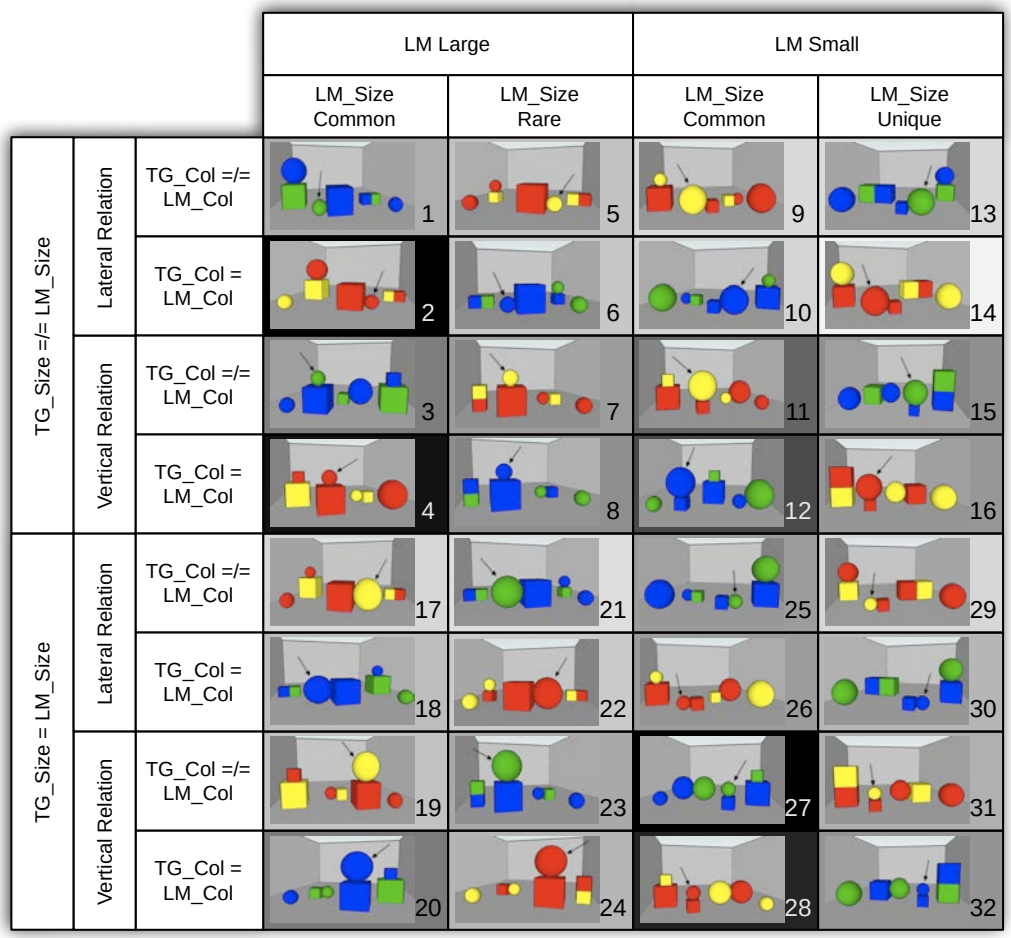

Figure 5.11: Scene Effects on the Use of Relations in GRE3D7. The darker the background of a cell, the higher the proportion of relational descriptions that were produced for this scene.

that, if the target is small, large landmarks are significantly more likely to be used than small ones (χ^2=9.35, df=1, $p \ll 0.01$), while if the target is large, there is a small but insignificant trend towards a preference for relations to different-sized (i.e. small) landmarks. This might explain why, overall, LM_Size did not have the expected effect.

It is also interesting to note that, in scenes with small targets and large landmarks where the landmark's size is common (Scenes 1–4), the colour similarity between target and landmark is the most influential factor, while in all other scenes, the type of relation has the highest impact.

5.5.3 Interim Summary

The GRE3D7 Corpus is considerably larger than GRE3D3 and was based on a more principled design of the stimulus scenes. It was designed to investigate the role that a landmark's size might play in the use of the spatial relation between the target and this landmark. The main outcomes of the analysis of the use of relations in the GRE3D7 Corpus are:

1. *The proportion of descriptions containing a relation was lower in GRE3D7 than in GRE3D3.* The most likely explanation is that the target–landmark relation in the GRE3D3 scenes was always unique as there was no other object pair in close proximity to each other. In the GRE3D7 scenes, on the other hand, there was always one other object pair that had the same spatial relation as the one holding between target and landmark. This suggests that the use of a relation is impacted not only by the visual properties of the landmark, but also by the discriminatory power of the type of the relation itself.

2. *The laziness effect was reduced due to filler scenes.* Although the laziness effect, which led people to reduce their use of relations over the course of the experiment, was not eliminated entirely in GRE3D7, it was somewhat reduced. It might be necessary to introduce an entirely different distraction task or to never collect more than one or two description from each participant at any point in time to fully overcome this problem.

3. *Just over half of the participants follow an exclusive strategy for the use of relations.* As was the case for GRE3D3, three broad participant-specific strategies can be identified with respect to the use of relations in GRE3D7. A large proportion of participants (135) opted to never use a relation, while a much smaller number of people (6) used a relation in all of their descriptions. The remaining 139 participants are responsible for the variation in the data, as they used a relation to describe the target in some but not all scenes.

4. *The target–landmark relation is used more often if it is vertical than if it is lateral.* This confirms previous psycholinguistic findings showing that humans prefer vertical relations and prepositions over horizontal, and in particular lateral, ones.

5. *If a landmark shares colour with the target it is more likely to be used in a referring expression.* This lends support to the hypothesis that visual similarity between target and landmark increases the likelihood of the relation

between them being used. However, similarity in terms of their size shows an opposite, but statistically not significant, trend. Therefore, if target–landmark similarity plays an important role in the use of spatial relations, colour seems to have a more decisive impact in this respect.

6. *Unexpectedly, landmarks with a common size were more often mentioned in referring expressions than those with a rare or unique size.* It is very likely that this is due to an unintended factor obscuring any effect the commonness of the landmark's size might have had. This factor was the length of the inherent MD for the target. In the majority of scenes with a common landmark size, the inherent MD was of length 3, while it was only of length 2 for scenes with a rare or unique landmark size. Targets with a longer inherent MD were described in terms of a relation more often than targets with a shorter inherent MD.

7. *The landmark's size itself had no effect on the use of relations, but an interaction between the factors* LM_Size *and* TG_Size=LM_Size *is evident:* small landmarks are used more often if TG_Size=LM_Size is true, while large landmarks are more likely to be included in a description if TG_Size=LM_Size is false. However, this might also be a direct effect of the target's size: in those situations where relations get used more often, the target is always small (either the landmark is small and the target's size is the same or the landmark is large and the target's size is different), while the target is large in those scenes with a lower probability for the use of a relation.

5.6 Variation in the Two Corpora

Each description contained in the GRE3D3 and GRE3D7 corpora can be characterised in terms of a **content pattern** defined by the presence or absence of each of the nine properties. For example, the description *the large blue ball* corresponds to the pattern ⟨tg_size, tg_colour, tg_type⟩. In this section, I examine the frequency of the different content patterns in the two corpora. Table 5.3 lists all patterns that occurred in at least one of the two corpora. Each content pattern was assigned an ID in the letter range A–ZK for easier reference. For each content pattern, the table lists the number of times this pattern occurs in each corpus and the proportion of the overall number of descriptions in the corpus that this pattern covers. Empty cells represent a zero count.

Interestingly, the smaller GRE3D3 Corpus contains 31 content patterns, four more than GRE3D7. The table reveals that two of the content patterns are over-

ID	Content Pattern	Example Description	Count		Proportion in %	
			GRE3D3	GRE3D7	GRE3D3	GRE3D7
A	⟨tg_col tg_loc tg_type⟩	blue cube in the front	6	0.95	29	0.65
B	⟨tg_col tg_loc tg_type rel lm_col lm_type⟩	the yellow ball on the right beside the red box	0	0	1	0.02
C	⟨tg_col tg_loc tg_type rel lm_size lm_col lm_type⟩	the blue ball on the ground by a big blue cube	0	0	2	0.04
D	⟨tg_col tg_type⟩	the yellow ball	172	27.30	1644	36.70
E	⟨tg_col tg_type rel lm_col lm_loc lm_type⟩	the yellow ball on top of the left hand red cube	17	2.70	0	0
F	⟨tg_col tg_type rel lm_col lm_type⟩	the green ball next to the blue cube	49	7.78	121	2.70
G	⟨tg_col tg_type rel lm_loc lm_type⟩	the red ball in front of the right-hand cube	5	0.79	0	0
H	⟨tg_col tg_type rel lm_size lm_col lm_loc lm_type⟩	the small green ball in front of the large blue cube on the left	0	0	1	0.02
I	⟨tg_col tg_type rel lm_size lm_col lm_type⟩	the green ball next to the big blue cube	12	1.90	46	1.03
J	⟨tg_col tg_type rel lm_size lm_loc lm_type⟩	the green circle in front of the left hand larger square	1	0.16	0	0
K	⟨tg_col tg_type rel lm_size lm_type⟩	the red cube on top of the big cube	2	0.32	6	0.13
L	⟨tg_col tg_type rel lm_type⟩	the yellow ball on top of the cube	2	0.32	14	0.31
M	⟨tg_loc tg_type⟩	rightmost ball	3	0.48	3	0.07
N	⟨tg_size tg_col tg_loc tg_type⟩	the little red ball on the right	2	0.32	39	0.87
O	⟨tg_size tg_col tg_loc tg_type rel lm_col lm_type⟩	the little red ball on the floor next to the red cube	0	0	1	0.02
P	⟨tg_size tg_col tg_loc tg_type rel lm_size lm_col lm_loc lm_type⟩	the small blue ball on the right on top of the large green cube on the right	3	0.48	0	0
Q	⟨tg_size tg_col tg_loc tg_type rel lm_size lm_col lm_type⟩	the small red ball on the left next to the small red cube	1	0.16	6	0.13
R	⟨tg_size tg_col tg_type⟩	small green cube	143	22.70	2145	47.88
S	⟨tg_size tg_col tg_type rel lm_col lm_loc lm_type⟩	the small green ball on top of the blue cube on the right	2	0.32	8	0.18
T	⟨tg_size tg_col tg_type rel lm_col lm_type⟩	the small green ball on top of the blue cube	31	4.92	93	2.08
U	⟨tg_size tg_col tg_type rel lm_loc lm_type⟩	the small green ball in front of the left hand cube	1	0.16	0	0
V	⟨tg_size tg_col tg_type rel lm_size lm_col lm_loc lm_type⟩	the small yellow ball on top of the large red cube on the right	6	0.95	0	0
W	⟨tg_size tg_col tg_type rel lm_size lm_col lm_type⟩	the little green ball to the left of the big blue box	30	4.76	238	5.31
X	⟨tg_size tg_col tg_type rel lm_size lm_type⟩	the big blue ball on top of the little cube	4	0.63	10	0.22
Y	⟨tg_size tg_col tg_type rel lm_type⟩	the small red ball on the box	2	0.32	19	0.42
Z	⟨tg_size tg_type⟩	the big ball	28	4.44	17	0.38
ZA	⟨tg_size tg_type rel lm_loc lm_type⟩	the little ball on top of the cube on the right	1	0.16	0	0
ZB	⟨tg_size tg_type rel lm_size lm_col lm_type⟩	the big ball on top of the big blue cube	0	0	1	0.02
ZC	⟨tg_size tg_type rel lm_size lm_loc lm_type⟩	the small ball next to the large cube on the left hand side	0	0	1	0.02
ZD	⟨tg_size tg_type rel lm_size lm_type⟩	the small cube in front of the big cube	1	0.16	8	0.18
ZE	⟨tg_size tg_type rel lm_type⟩	the small ball on top of the cube	11	1.75	3	0.07
ZF	⟨tg_type⟩	the ball	52	8.25	3	0.07
ZG	⟨tg_type rel lm_col lm_type⟩	the ball on top of the red cube	5	0.79	3	0.07
ZH	⟨tg_type rel lm_loc lm_type⟩	the ball on top of the cube on the right	10	1.59	0	0
ZI	⟨tg_type rel lm_size lm_col lm_type⟩	the ball on top of the big red cube	1	0.16	0	0
ZJ	⟨tg_type rel lm_size lm_type⟩	the ball on top of the large cube	5	0.79	0	0
ZK	⟨tg_type rel lm_type⟩	the ball on top of the cube	22	3.49	18	0.40

Table 5.3: The 37 different content patterns that occur in the GRE3D3 and GRE3D7 corpora.

whelmingly common in both corpora. These are the patterns D and R, shown in Examples (5.10) and (5.11). These two patterns combined cover 84.6% of GRE3D7 and 50.0% of GRE3D3. It is not surprising per se that these content patterns are common, as one of them is the inherent MD for every stimulus scene in GRE3D7 and for all GRE3D3 scenes which are based on Schemata D and E. However, two things are surprising about the frequent use of these patterns: Firstly, they cover a much larger portion of GRE3D7 than the four inherent MDs of GRE3D3 taken together. These four (patterns D, R, Z and ZF; see Examples (5.10) to (5.13)) only cover 62.7% of GRE3D3. And secondly, they are much more frequent in GRE3D3 than the other two inherent MDs for this corpus, ZF for scenes based on Schemata A and B (8.3%) and Z for scenes based on Schema C (4.44%).

(5.10) D: ⟨tg_colour, tg_type⟩

(5.11) R: ⟨tg_size, tg_colour, tg_type⟩

(5.12) Z: ⟨tg_size, tg_type⟩

(5.13) ZF: ⟨tg_type⟩

The second of these two observations is easily explained by the oft-observed primacy of colour as the most common property: Patterns D and R both contain the attribute tg_colour while Z and ZF do not. However, the fact that such a large portion of GRE3D7 is covered by just two content patterns remains interesting. It means that this data set, which is much larger, not only contains fewer different content patterns, it is also spread less evenly across those patterns than the GRE3D3 corpus. This difference is reflected in the difference in variance of the pattern frequency distributions of the two corpora: GRE3D3's variance $\sigma^2_{GRE3D3} = 37.66$, while GRE3D7's variance $\sigma^2_{GRE3D7} = 127.71$, which is a statistically highly significant difference ($F(26, 30) = 3.39$ with probability $P(F(24, 30) \geq 2.47) = 0.01$). This means that the distribution of patterns in GRE3D3 is closer to a uniform distribution than that of GRE3D7. In GRE3D7, with its high variance, each pattern gets used either much more or much less than the mean, while in GRE3D3, with its lower variance, the frequency of use of each pattern is closer to the mean than in GRE3D7.

Another way to quantify the difference in variance between the two corpora is to look at the average entropy of the patterns contained in them. The average entropy of a content pattern in GRE3D3 is 3.1463 bits, while the patterns in GRE3D7 have an average entropy of only 1.942 bits, which means that the surprisal factor of the patterns in GRE3D3 is much higher, making them harder to predict.[9]

[9] I am not aware of a statistical test that can check for significance between two entropy values.

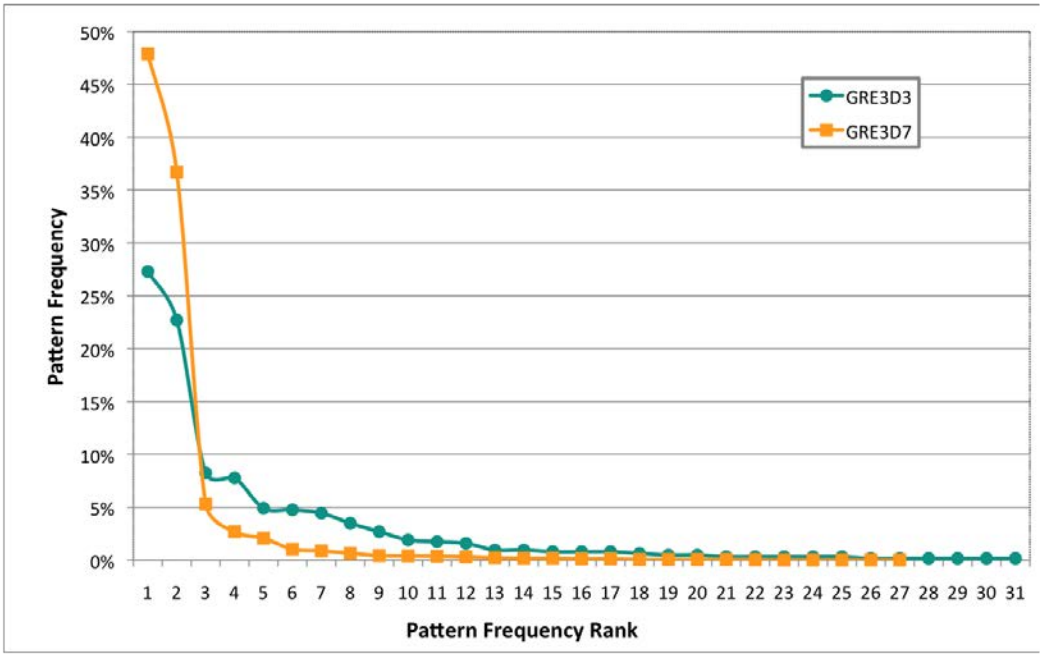

Figure 5.12: Percentage of Pattern Use by Frequency. The curves show the decrease in frequency for the content patterns in GRE3D3 and GRE3D7 from most frequent to least frequent patterns.

Both variance and entropy intuitively compute the level of 'flatness' of a distribution curve. So, comparing variance or entropy values is not sensitive to different shapes of distribution curves. Figure 5.12 shows the distribution curves for the pattern usage from most frequent to least frequent for both GRE3D7 and GRE3D3. It visualises a comparison of the second columns for each corpus in Table 5.3. As we can see, the curve for GRE3D7 starts at a much higher point for its most popular patterns than that for GRE3D3 and also drops faster, which makes the two curves *look* different.

The Kolmogorov-Smirnov test is a non-parametric test that can be used to determine whether there is a statistically significant difference between two distribution curves, such as those shown in Figure 5.12. It tests for difference in the curves' medians, dispersion and skewness. In this case, the *Kolmogorov-Smirnov* Z value is 1.701, which rejects the null hypothesis with a probability of error $p = 0.006$.

A possible explanation for the lower variability in GRE3D7 could be that, despite the longer duration of the data gathering experiment, people got less bored. This could be due to the variety of filler scenes that they were shown between the stimulus scenes as well as to the slightly more complex nature of the stimulus

scenes themselves. Being less bored with the task might have made participants less likely to try to find ways to make it more exciting by varying the descriptions they gave. Another, related, factor that could be at play is that due to the filler scenes people did not notice how similar the stimuli were or that they were giving very similar descriptions for all of them.

5.7 Conclusions

In this chapter, I have presented the design, collection and analysis of two corpora of distinguishing descriptions. Both were designed to allow investigations into the circumstances under which people use spatial relations in referring expressions, as the use of spatial relations has proven to be one of the major challenges for existing REG algorithms, as discussed in Chapter 4.

The stimuli for both corpora consisted of simple 3D scenes containing either 3 or 7 geometric objects which differed in colour, size and type. In each scene one object was marked as the intended target referent and one landmark object was located very close to the target to encourage participants to use the relation to this landmark, if they were going to use a landmark at all. The target could always be fully distinguished from all other objects in the scene by its colour, size and shape only; spatial relations or other locational information were never necessary.

The following is a list of the conclusions I draw from the analyses presented in this chapter:

1. *People use relations, even when this is not necessary.* This is one of the foremost conclusions of this chapter and it is true in both corpora. However, the use of relations was much lower in GRE3D7 than in GRE3D3. This strongly suggests that the discriminatory power of the spatial relation itself is of importance. In the GRE3D7 scenes there was always a second object pair with the same spatial relation as the target–landmark pair, while in the GRE3D3 scenes the target and landmark were the only pair of objects in close proximity to each other.

2. *Much depends on the preferences of the participants.* In both corpora, I found three broad strategies with respect to the use of spatial relations: many people chose never to use a relation, some used a relation in all their descriptions, and the remaining participants did not follow such an exclusive strategy but rather varied their behaviour between the stimulus scenes.

3. *A strong temporal effect shapes the use of relations.* Those participants who did not follow an exclusive strategy were much more likely to use a spa-

tial relation at the beginning of the experiment than towards the end of it. Without such a temporal effect, people's natural likelihood to use a spatial relation in a one-off description can therefore be expected to be much higher than the average use of relations observed in the corpora.

4. *Spatial relations in the vertical axis are preferred over those in the horizontal axes.* This is in accordance with previous psycholinguistic findings.

5. *The length of the inherent minimal description impacts on the use of spatial relations.* One underlying hypothesis examined in this chapter was that for targets which are easy to describe using only inherent visual properties, a spatial relation would be less likely, while targets for which many inherent visual properties have to be listed would be more likely to be described in terms of a spatial relation to a landmark. In GRE3D3 I found no reliable effect of the length of the target's inherent minimal description on the use of relations, possibly due to too many overlapping conditions in this corpus. However, in GRE3D7, the length of the inherent minimal description turned out to be an unintended confounding factor with a significant effect: targets with an inherent minimal description of length 3 were twice as likely to be described in terms of a spatial relation than targets with an inherent minimal description of length 2.

6. *Visually salient landmarks are more likely to be used in a description.* Based on the literature on visual salience and attention (Pashler, 1998; Yantis and Egeth, 1999; Caduff and Timpf, 2008), I adopted a definition of visual salience based on attribute difference: rare attribute values make an object stand out visually from its physical context. In GRE3D3, I tested whether the landmark's similarity to each of the other two objects in the scene had an influence on the use of relations, as well as whether there was an effect of the overall distinctness of the landmark from the other two objects. I found that a unique landmark is used reliably more often in a referring expression than one that is similar to at least one of the other two objects. I also found that a relation is less likely to be used if the landmark resembles the distractor than if it resembles the target object. In GRE3D7, I tested whether landmarks with a common size are more likely to be used in referring expressions than landmarks with a rare or even unique size. However, any effect of this factor was overshadowed by the effect of the length of the target's inherent minimal description.

7. *Target and landmark size in interaction influence the use of relations.* Two hypotheses underlying the design of the GRE3D7 Corpus were (1) that large

landmarks would be more visually salient and therefore used more, and (2) that the difference between the target's and the landmark's size might be of more importance than the difference between the landmark's size and that of the other objects. Neither of these two factors showed a reliable independent effect. However, there was a significant effect of these two factors in interaction with each other, which indicates that the size of the target might be of more importance than I had expected: small targets are more likely to be described in terms of a spatial relation than large ones. In interaction with the target's size, the landmark's size has a limited effect: if the target is small, large landmarks are significantly more likely to be used than small ones, and if the target is large, there is a statistically not significant trend for small landmarks to be used more often.

8. *The larger GRE3D7 Corpus is less varied than GRE3D3.* An analysis of the content patterns used in the two corpora revealed that, interestingly, there was less variation in the GRE3D7 Corpus, despite its being much larger than GRE3D3. The nine properties (tg_size, tg_colour, tg_location, tg_type, relation, lm_size, lm_colour, lm_location and lm_type) were combined into 31 different content patterns in GRE3D3, but GRE3D7 contains only 27 content patterns. Furthermore, the distribution across the different patterns is more uniform in GRE3D3.

The design of the corpora presented in this chapter was of course limited in a number of regards. For example, the objects used in the stimulus scenes are not realistic objects, but rather abstract geometric shapes, and the number of objects in each scene as well as their visual properties were intentionally kept low. However, using such controlled settings is the only way in which we can hope to be able to make sense of people's reference behaviour. While the analysis of the corpora presented in this chapter cannot explain every nuance of reference 'in the wild', it does bring us a step closer to a full understanding of the factors that influence the semantic content that people choose to include in referring expressions.

A final important contribution that this chapter has to make to the field of referring expression generation is the provision of two new, large corpora of distinguishing descriptions. They are particularly useful as they contain many descriptions for each different stimulus item capturing at least some of the variety in choices that speakers make when they work out how to refer to an object in a given situation. Hopefully, the existence and accessibility of such corpora will encourage future corpus-based work with the aim of increasing the naturalness of the output of REG algorithms.

Chapter 6

Corpus-Based Modelling of REG

In Chapter 5, I described two corpora of human generated referring expressions which pick out objects in small visual scenes. The design of these scenes was based on the hypothesis that the visual salience of the target and the potential landmark objects impacts on the use of spatial relations. In this chapter, I will take a slightly different view on visual salience and apply it to the design of machine learning features, in order to characterise the overall referring behaviour displayed in these corpora using a simple statistical model. This characterisation will help answer two questions: first, what are the relevant factors that affect whether a human is likely to include a particular property in a referring expression? Second, how do these factors interact?

A natural way to tackle the first of these questions is to train a statistical classifier using characteristics of the referential context as features. I can then determine whether the chosen features indeed have an effect on the content of the referring expressions in my corpora by observing how accurately the model produced by the classifier predicts the data. A decision tree learner is an obvious candidate to help answer the second question of how the features interact, because decision trees are easy to inspect and interpret.

Based on the evidence from the previous two chapters, I do not expect that it is possible to learn one general model that is able to characterise the referring behaviour of all participants. The analysis of the human data in the GRE3D3 and GRE3D7 corpora showed that much of the variation in semantic content of referring expressions cannot be explained by the features of the scenes alone. I use two techniques to deal with this problem: firstly, I train decision trees that predict only one attribute at a time and establish that at this level more commonality in people's behaviour can be found. Secondly, I train models that take into account the identity of the participant. The attribute-specific trees of one participant

can then be combined into a **speaker profile** for this person, and the individual component trees of this profile can be compared to other participants' strategies, even if the overall speaker profile is unique in the corpus.

Section 6.1 lays out the framework for the experiments by describing the prediction classes and the features available to the learner. In Section 6.2, I present decision trees that attempt to predict the complete content of referring expressions. Following this, I examine the performance of decision trees modelling the use of individual object attributes in Section 6.3. In Section 6.4, I examine how well the models learned on each corpus perform on the other corpus. This delivers some interesting insights into the similarities and differences between them. Section 6.5 examines the extent to which speaker-specific behaviour patterns are responsible for the variation in the data by, first, allowing the participants' identity as a decision feature in the decision trees and, second, training individual trees for each speaker which can be compared to each other to find commonalities between speakers.

6.1 Setting Up the Experimental Framework

6.1.1 The Prediction Classes

As I described in Chapter 5, the participants that contributed to the two corpora under examination were shown simple scenes and asked to produce a referring expression to pick out a particular object. By design, the scenes were limited in detail to keep the number of properties that would be included in the referring expressions low. The experiments I report in this chapter will characterise the referring expressions according to whether a particular property was included in the referring expression or not. This section describes the set of properties whose use I attempt to model.

The analysis of the corpora in Chapter 5 showed that people mostly used the properties I anticipated based on the way the scenes were constructed. Those properties were the type, colour and size of the target object, the relation to the closest landmark object and again the type, colour and size of this landmark. In a few cases (9.84% in the GRE3D3 corpus and 2.03% in GRE3D7) a participant included a non-relational locational property for either the target or the landmark object. Description (6.1) is an example from GRE3D3 in which the landmark object is specified by using its location. In Description (6.2), the location of the target itself is used.

(6.1) the ball in front of the rightmost cube (GRE3D3; Scene 16)

(6.2) the large red ball in the middle of the scene (GRE3D7; Scene 14)

tg_type: the target object's type

tg_colour: the target object's colour

tg_size: the target object's size

tg_location: the target object's location

relation: the spatial relation between the target and the landmark

lm_type: the landmark object's type

lm_colour: the landmark object's colour

lm_size: the landmark object's size

lm_location: the landmark object's location

Figure 6.1: The nine constituent properties used in the referring expressions.

With ten instances in GRE3D3 and six in GRE3D7, relations to another object than the landmark were extremely uncommon. They only occurred in descriptions that also mentioned the landmark object and usually involved the ternary relation *between* as in Description (6.3).

(6.3) The green ball between the two cubes. (GRE3D7; Scene 1)

These mentions of a third object are excluded from the data experiments in this chapter: I do not train trees to predict whether to use a third object in a referring expression as there are too few instances of a third object being mentioned. This leaves us with the nine properties shown in Figure 6.1. As shown in Chapter 5, these properties were combined into what I call a **content pattern** in 31 different ways in the GRE3D3 corpus and in 27 different ways in the GRE3D7 corpus, with an overlap of 21 patterns. In the experiments described below, I first train decision trees to predict which of these content patterns should be used in a given situation. Following this, I train trees that make more fine-grained decisions about the inclusion of each individual property in the list above, with the exception of tg_type and lm_type. tg_type was used in every description and lm_type in every relational description. Therefore, a simple rule which always includes them suffices to model the reference behaviour regarding these two.

6.1.2 Features to Learn From

In this section, I describe the features that I use in the machine learning experiments below. The features were chosen based on the expectation that they would

have an impact on people's choice of attributes for a referring expression. The rationale behind this expectation will become clear over the course of this section. There are two different types of features: **scene-independent** features, which describe aspects of the external situation in which the description was produced and do not change between scenes; and **scene characteristics**, which describe visual attributes of the scene and the objects in it.

In the case of the GRE3D3 and GRE3D7 corpora, scene characteristics are the main source of features that I can draw on for a machine learning exercise. The only scene-independent feature I will use is the Participant_ID, which identifies the participant in the data gathering experiment in which the corpus was collected. This can only be used as a feature when learning and testing on the same corpus, as the participants were different for the two data gathering experiments. If people have idiosyncratic ways of referring to objects, information about who produced a certain referring expression should increase the chances of correctly predicting the content of the referring expression. For each prediction class, I trained two classifiers: one including the Participant_ID and one not including it. The difference between the results of these will give us an idea of how participant-dependent any variation in the data is. These results are reported in Section 6.5.1.

The remaining, scene-dependent, features can be further categorised into **direct object properties**, which simply record the attribute value of a certain object in the scene, and **comparative features**, which compare the attribute values of on object to those of the other objects. In the following, I discuss these two different types in more detail.

Direct Object Properties as Features

The simplest way to encode what a scene looks like is to create features that record which value each object takes for each of its attributes and which spatial relations hold between all of the objects in the scene. Such a basic level feature set would have the advantage that it would implicitly contain information not only about, for example, how many objects share their size and how many share colour, but also about how many objects share both size *and* colour. However, defining such a feature set would involve a lot more difficulty than is immediately apparent.

In order to be able to define these features it would be necessary to first identify in an unambiguous way which object in one stimulus scene corresponds to which object in another one. In the GRE3D3 scenes this is relatively unproblematic as there are only three objects: the target, the landmark closest to the target and one

distractor.[1] Similarly, the target and closest landmark can be clearly identified in each stimulus scene of the GRE3D7 corpus.

Deciding for the remaining five objects in the GRE3D7 stimuli how they correspond to each other across scenes, on the other hand, is far from trivial, despite a certain degree of common structure. For the cross-corpus comparison in Section 6.4, I would also need to define an object in each GRE3D7 scene that corresponds to the distractor object in each GRE3D3 scene, and find an adequate way to deal with the empty feature values that each GRE3D3 instance would have for the remaining four GRE3D7 objects. Neither of these two problems has a satisfactory solution.

I therefore only define features for the attributes of the target and landmark objects, which are clearly identifiable in every scene in both corpora. The possible candidate attributes which can be turned into features are the type, size, colour and location of the target and landmark as well as the relation between them.

I did not include the location in the scene of the target and landmark objects as features because the stimulus scenes for both corpora were constructed in such a way as to keep this variable as constant as possible. In the GRE3D3 scenes the target and landmark were always to one side of the third object and, as mentioned in Chapter 5, which side they were on had no impact on the content of the referring expressions used. In the GRE3D7 scenes the target–landmark pair was always positioned as close to the centre of the scene as possible, which means that their location is not a feature based on which it would be possible to distinguish scenes from each other. The values of features recording the location of target and landmark would be the same for all scenes.

This leaves us with the values of type, colour and size of target and landmark as possible absolute attribute features as well as the relation between these two objects. Of these, type and colour were also not included as features. The reasoning underlying this decision is that the different values these attributes can take, cube or ball for type and blue, green, red or yellow for colour, are not different from each other in visual salience per se. For example, green is only more salient than blue in a given scene if it is very rare and blue is very common. So, it is the number of times an object shares these properties with other objects that might make a difference, rather than the attribute values themselves.

The actual value of the size of an object, on the other hand, might very well have an impact on the object's salience: a large object takes up much more real estate in a scene and stands out more than a small object. Of course, the values large and small for the attribute size are not actually absolute, but rather they are relative

[1] See Section 5.2.1 regarding the naming of these objects.

values that are only defined in comparison to the size of other objects. However, it is exactly this relativity, the fact that a large object is large in comparison to the other objects in the scene and that it stands out more than other objects, that leads me to give size a special role as a feature for the learning experiments.

The absolute scene characteristics are then TG_Size, LM_Size and Relation_Type, the type of spatial relation between the target and landmark objects (see Table 6.1 for a full list of all features). The values for Relation_Type are horizontal and vertical because the target object was placed either on top of or in a horizontal relation to the landmark object: in front of the landmark in the GRE3D3 stimuli and left or right of the landmark in the GRE3D7 stimuli. Relation_Type is included because the data analysis in Chapter 5 showed that participants in both data gathering experiments were more likely to include the relation when it was a vertical one than when it was a horizontal one.

Comparative Scene Characteristics

Instead of recording the properties of objects directly, comparative scene characteristics say something about a certain property of one object compared to that of one or more other objects. These features are based on an approach to approximating visual salience. The assumption is that an attribute is more likely to be included in a referring expression if its value is uncommon in the scene, not only because that makes it more useful for distinguishing the target from the other objects, but also because its rareness makes it stand out visually. Similarly, uncommon properties of the landmark object, making it more useful for distinguishing the target and also more visually salient, might increase the likelihood of a relation to this landmark being included.

The fact that a property or a relation to a landmark is useful for distinguishing the target from other objects (i.e its discriminatory power) needs to be computed by comparing it to all other objects and counting how many are ruled out as distractors if this property or landmark is included in the referring expression. At the same time, however, the accompanying high visual salience is more directly coupled to the visual perception of the scene and omits the step of resource-intensive computation, which might result in the inclusion of a property or landmark.

This lead me to design features that record how many objects the target shares its properties with in order to capture their usefulness and their visual salience. I also introduced the equivalent features for the landmark, based on the reasoning that the fewer objects share the landmark's properties, the more visually salient it is and the more inclined a person might be to use the relation to this landmark

	Feature	Explanation	Values
direct	TG_Size	size of the target object	small, large
	LM_Size	size of the landmark object	small, large
	Relation_Type	type of relation between target and landmark	horizontal, vertical
comparative	Num_TG_Size	number of objects of same size as the target	numeric
	Num_LM_Size	number of objects of same size as landmark	numeric
	TG_LM_Same_Size	target and landmark share size	Boolean
	Num_TG_Col	number of objects of same colour as target	numeric
	Num_LM_Col	number of objects of same colour as landmark	numeric
	TG_LM_Same_Col	target and landmark share colour	Boolean
	Num_TG_Type	number of objects of same type as target	numeric
	Num_LM_Type	number of objects of same type as landmark	numeric
	TG_LM_Same_Type	target and landmark share type	Boolean
	Participant_ID	ID number of the description giver	alphanumeric

Table 6.1: The features and their value formats.

as well as its rare properties in a referring expression.

The centre part of Table 6.1 lists the comparative features I use in the experiments below. The features whose names start with Num_ record how many objects in a given scene are of the same type as the target or landmark object. Those starting with TG_LM_ only compare the target's properties to those of the landmark. The TG_LM_ features are included under the assumption that these two objects are in or closest to the visual focus of the participants, which makes it likely that their respective properties have an especially high impact on the content of referring expressions in my corpora.

Of course, it would be possible to extend the list of features and include the same comparisons for all objects in a scene, not only the target and the landmark: I could try to record for each object how many other objects it shares its colour, size and type with. There are a number of reasons that speak against taking this path. Firstly, this would mean a considerable amount of redundancy in the information provided in the feature set: the fact that four objects share colour in a scene would then be expressed four times, once for each of these objects. And secondly, in order to create a feature recording the different values for this one object in the different scenes, we would need a way to unambiguously determine which object in one scene corresponds to which object in another scene, which, as I argued earlier in this section, is at least in the GRE3D7 scenes by no means an obvious decision.

6.1.3 Feature Values

The last column of Table 6.1 shows what type of values each of the machine learning features takes. The features comparing target and landmark properties to each other, such as TG_LM_Same_Size, take Boolean values recording whether it is TRUE or FALSE that target and landmark share the property. The direct object features TG_Size and LM_Size take their values from the set {small, large}, and Relation_Rype takes its values from {horizontal, vertical}. The Participant_ID is an alphanumeric symbol consisting of a combination of the name of the corpus to which this participant contributed and a running number, e.g. GRE3D3-63 is participant 63 from the GRE3D3 corpus. The features recording the number of objects the target or landmark share a particular property with, such as Num_TG_Type, are numeric.

In order to be able to train a decision tree on one of the two corpora and test its performance on the other corpus, the features used by the machine learners need to have comparable values for both corpora. In the case of alphanumeric features this means that the values have to come from the same set of possible strings. For example, it would not be possible to use the feature Relation_Type if the value set was {frontal, vertical} for the GRE3D3 corpus (where the target was either on top of or in front of the landmark) and {lateral, vertical} for the GRE3D7 corpus (where the target was either on top of or to one side of the landmark). In order to make this feature comparable across the two corpora it is necessary to collapse the values frontal and lateral into a higher-level value that subsumes both: horizontal.

A more subtle case concerns the numeric features recording the number of objects sharing properties with the target and landmark. Although there is technically no problem with comparing the exact counts represented by these values to each other, conceptually they mean quite different things in the two corpora: if the target shares its type with two objects in the a scene from GRE3D3, that means that all objects in the scene are of the same type. However, the same count in a GRE3D7 scene amounts to less than half of the objects sharing the target's type. Additionally, making a distinction between four or five objects sharing a property seems to be giving too much importance to the exact counts over the underlying rationale for the use of these features: the attempt to capture a more vague notion of visual salience. It seems unlikely that it is the absolute number of objects the target or landmark share properties with which makes a large difference to the choice of content for a referring expression, but rather the *proportion* out of all objects in the scene.

As a solution to this problem, I translate the numerical values in the two corpora into a range of slightly more vague concepts such as none, half, most and

	number of objects sharing the property with the TG or LM				
GRE3D7	0	1 2	3	4 5	6
GRE3D3	0		1		2
general scale	none	few	half	most	all
joint coding	0	1	2	3	4

Table 6.2: Joint coding for both corpora of the numeric features. The coding is based on the number of objects that share a property with the target or landmark object.

so on and then re-code these concepts on a new numerical scale that is the same for both corpora. This way of coding the numerical values has three advantages: Firstly, it guarantees that important values such as all and none are represented the same way independently of how many objects are in each scene. Secondly, all instances can be classified, even if the test corpus has more possible values than the training corpus. A tree splitting on a categorical feature would not be able to classify test instances which have a value for this feature that did not occur in the training set. Thirdly, retaining the numerical nature of the comparative features allows the learner to introduce elegant two-way splits using inequalities rather than having to make complex multi-way splits along several values.

Table 6.2 shows the consolidated coding for the numeric features in the two corpora. I introduce a general scale which records whether a property is shared with none, few, half, most or all objects in the scene. The table also shows how the possible values on the two corpora are mapped onto a new, joint, numerical scale ranging from 0 to 4.

6.1.4 Decision Tree Classifiers

I used the Weka workbench (Witten and Frank, 2005) for the machine learning experiments presented below. Weka provides a large number of machine learning schemes including J48, which implements the classic decision tree algorithm C4.5 by Quinlan (1993). The main advantage of decision tree learners over many other machine learning schemes is the ease with which their output can be understood and interpreted and the way rules they produce can be used as heuristics in larger NLG systems.

The C4.5 algorithm builds a decision tree from the root up. Starting with the complete set of all data instances, it recursively selects one feature for each node; then the data instances at this node are split into branches according to this feature. The choice of feature is based on the information gain, measured in bits,

that a split causes for the overall tree. At each decision point the feature resulting in the highest information gain is chosen. If the feature has nominal values, one branch is created for each value of the feature. Continuous numerical features result in a binary split with one branch for all instances with a smaller value than a certain threshold and one branch for those with a larger value. The threshold is again determined by finding the highest possible information gain.

Once the tree is built, a postpruning step reduces the complexity of the tree in an attempt to avoid overfitting to the training data. Pruning can increase the performance of a decision tree on unseen test data, but often decreases the performance on the training data which the original unpruned tree was based on. The details of the pruning algorithms used by C4.5 are described in (Witten and Frank, 2005, pp. 192–198). Because I am not only interested in seeing whether the trees learned on my data generalise well to unseen test data, but also which features are actually used in the decision trees, I compute the results for both pruned and unpruned trees in all experiments.

Also because I am not just interested in producing decision trees that will perform well given a new problem (a new object to be described in a similar scene), but also in finding trees that represent data sets as faithfully as possible, I change the standard value of two minimum instances per leaf to one. This gives the learner the chance to split on a feature even if some of the resulting branches contain only one instance.

As a baseline for the experiments in this chapter, I use the majority class learner implemented in Weka's zeroR classifier. For example, if we want to learn whether to include a relation in a referring expression and most descriptions in the training corpus did not contain a relation, the majority class model will never include a relation for any instances of the test set either.

6.2 Modelling the Use of Complete Content Patterns

In Chapter 5 we saw that each description in the two corpora can be characterised in terms of its content pattern, the set of properties it contains.[2] In the first of the machine learning experiments I report in this chapter, I trained decision trees that predict exactly which of the content patterns a description for a given object in a given stimulus scene should follow. The tree learner was able to choose from the twelve features based on characteristics of the scene, excluding Participant_ID. As we will see, the high variation between participants means that the pattern to

[2] In Table 5.3, I provided a list of all the content patterns together with their absolute and proportional frequencies in the two corpora.

training corpus	test method	baseline	pruned tree	unpruned tree
GRE3D3	10 fold X	27.30%	**46.51%**	**46.19%**
	training set	27.30%	**46.51%**	**46.51%**
GRE3D7	10 fold X	47.88%	**64.71%**	**64.93%**
	training set	47.88%	**64.93%**	**64.93%**

Table 6.3: Accuracy for the trees characterising the use of whole content patterns. Bold values are statistically significantly different from the baseline at $p<0.01$ using the χ^2 test.

D: ⟨ tg_col, tg_type ⟩
R: ⟨ tg_col, tg_size, tg_type ⟩

Figure 6.2: Content patterns D and R.

use in any given instance is very hard to predict.

As mentioned before, I am interested in both the pruned and the unpruned trees. In Table 6.3 we see for each corpus the prediction accuracy that was achieved by the trees. The prediction accuracy represents the percentage of test instances for which the learned tree predicted exactly the same content pattern as found in the corpus. The results table shows that, while the decision trees significantly outperformed the majority class baseline, they still achieve rather low accuracy. The small differences between the performance of the pruned and the unpruned trees as well as between testing on the complete training data versus using ten-fold cross-validation show that the trees are not overfitted to the training data to a large degree, so they should have a good chance of achieving the same results on new, unseen data gathered in the same settings. The results for GRE3D7 are much better than those for GRE3D3. This is not very surprising, if we take into account the difference in variance between the two corpora which was discussed in Section 5.6. The entropy of the pattern distribution in GRE3D3 is much higher than that of GRE3D7, which makes predicting GRE3D3's patterns much harder.

Let's have a closer look at the actual predictions and trees. The majority class baseline rule predicts, of course, the most common content pattern for each corpus.[3] This is pattern D for GRE3D3, consisting of the target's colour and type, and pattern R for GRE3D7 which additionally includes the target's size. For ease of reference, patterns D and R are shown in Figure 6.2.

[3]Note that the rules learned for the ten folds in cross-validation might be different from each other, however the training software only returns the tree learned on the complete data set.

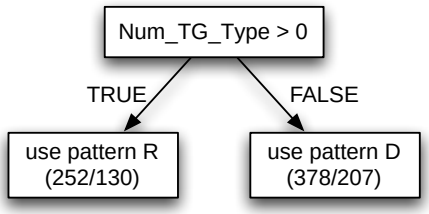

Figure 6.3: The GRE3D3 tree predicting the full content pattern.

It is not possible to inspect each of the trees trained on the different folds in the ten-fold cross-validations, as the classification software only returns the tree trained on the full data set, no matter which evaluation method is used. As it turns out, the pruning step did not make any difference to the tree trained on the complete GRE3D3 corpus, although the slight difference in the accuracy shows that there must have been a small change in at least one of the folds' tree. The tree trained on the full GRE3D3 set is displayed in Figure 6.3. It only uses one rule, splitting the data instances into two sets: for scenes in which the target's type is unique in the scene, the most common pattern, D, is predicted and for scenes in which the target shares its type with another object, pattern R is predicted, which is the second most common in the GRE3D3 data. The numbers in parentheses at each leaf show how many data instances were classified into this leaf and how many of these were misclassified. The numbers show that the classification error was roughly the same for both leaves in this tree.

The pruned and unpruned trees trained on the GRE3D7 data differed slightly from each other because the pruning step removed one binary split from the tree thereby reducing the number of leaves from 9 to 8. The pruned tree is shown in Figure 6.4. What we see is that, despite the much higher number of leaves than in the GRE3D3 tree, again only two different content patterns are predicted, and again these two patterns are D and R, the ones that were used most frequently in both corpora, as we saw in Table 5.3.

The tree uses six of the twelve possible features. Most interesting is the use of the feature Relation_Type, which specifies whether the relation between the target and the landmark was vertical or horizontal. The only difference between the two predicted content patterns is the presence or absence of the target's size. It seems slightly odd that the type of relation should have an impact on the use of size. Drilling down into the tree, we can see that the Relation_Type feature is used to split between Scene 17 and Scene 19, which contain exactly the same objects, three

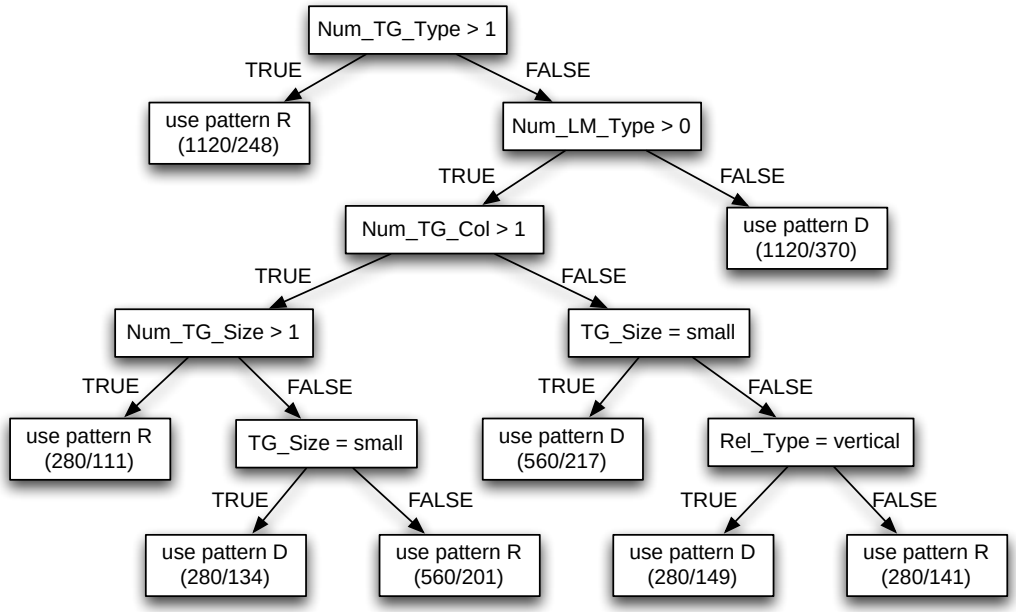

Figure 6.4: The pruned GRE3D7 tree predicting the full content pattern.

of which are large, including the target and landmark objects.

The only difference in terms of the intended design features is the type of relation. However, in Scene 17, the one for which the tree in Figure 6.4 includes tg_size, the third large object is located slightly further in the background, which makes it appear a bit smaller than in Scene 19, where it is almost at the same level as the target and landmark. In Scene 17 the property value large might therefore appear to be more distinguishing than in Scene 19. As we have seen in Chapter 5, the use of tg_size, as opposed to any of the other properties, is highly dependent on its usefulness to distinguish the target from any distractors. Possibly, this fact allowed the decision tree learner to capitalise on an unintended difference in the appearance between two scenes.

As we saw above, the accuracy of the decision trees for neither of the two data sets was particularly high, which could be due to a few different factors. It might be that the variation in the data is simply too random to be predicted accurately at all, or it might be that the features I chose to characterise the scenes do not capture the important differences between the data instances. Another possibility is a data sparseness problem: the corpora are just not big enough and the number of different possible content patterns are too many for the machine learner to be able to find robust associations between the features and the prediction classes.

A confounding factor is the discreteness of the set of prediction classes. This discreteness masks the fact that many of the content patterns are quite similar to each other with large overlaps in properties. The evaluation in terms of prediction accuracy does not take similarities between prediction classes into account. If a tree predicts a different class from the one in the data set, this is counted as an incorrect instance, regardless of whether the predicted content pattern differs only in one property from the content pattern in the data set or whether the two patterns have no overlap at all. No partial credit is given to predictions that are at least close to a human-produced content pattern for the same instance.

6.3 Modelling the Use of Individual Attributes

To overcome the problem of the actual content patterns being not all distinct, but rather overlapping in many cases, while the learning classes representing them are necessarily treated as discrete entities, I used decision trees that predict for each individual attribute whether it should be included in the referring expression of the target in a given scene or not. This reduces the number of prediction classes to two down from 31 in the case of GRE3D3 and 27 in GRE3D7, also hopefully alleviating any data sparseness problem that might exist.

Using decision trees in this way permits more fine-grained insights into the way the participants chose the content for the referring expressions they provided. While there was only a limited amount of commonality at the level of full content patterns, it might be the case that the use of individual attributes shows more regularities in terms of the scene characteristics I defined in Section 6.1.2. Splitting up the content patterns also makes it possible to give partial credit to this data-based approach for getting at least some of the attributes right, rather than having to get the complete content right or be penalised as if there was no overlap with the human-produced data at all.

The attributes for which I trained decision trees are those described in Section 6.1.1. In these experiments I treat target attributes and landmark attributes in a very similar way. This is only possible because the vast majority of referring expressions in the corpora only contained references to the target and the *intended* landmark, and therefore no decision has to be made regarding which object to include via a spatial relation. However, the trees characterising the use of the landmark attributes are trained on the relational data only in order to get meaningful results for these trees. This is discussed further in Section 6.3.2. In the following, I first present the results for the decision trees modelling the use of the target's attributes.

	training corpus	test method	baseline	pruned tree	unpruned tree
target colour	GRE3D3	10 fold X	77.94%	77.94%	77.94%
		training set	77.94%	77.94%	77.94%
	GRE3D7	10 fold X	98.72%	98.72%	98.72%
		training set	98.72%	98.72%	98.72%
target size	GRE3D3	10 fold X	57.78%	**90.48%**	**90.48%**
		training set	57.78%	**90.48%**	**90.48%**
	GRE3D7	10 fold X	57.75%	**73.95%**	**73.95%**
		training set	57.75%	**73.95%**	**73.95%**
target location	GRE3D3	10 fold X	97.62%	97.62%	97.62%
		training set	97.62%	97.62%	97.62%
	GRE3D7	10 fold X	98.19%	98.19%	98.19%
		training set	98.19%	98.19%	98.19%
relation	GRE3D3	10 fold X	64.44%	64.44%	66.34%
		training set	64.44%	65.87%	66.34%
	GRE3D7	10 fold X	86.61%	86.61%	86.61%
		training set	86.61%	86.61%	86.61%

Table 6.4: Accuracy for the trees characterising the use of each target attribute. Bold values are statistically significantly different from the baseline at $p<0.01$ using the χ^2 test.

6.3.1 The Target's Attributes

Table 6.4 shows the accuracy of the decision trees for each attribute of the target including the relation, and Table 6.5 shows the sizes of these trees. The size of a tree is measured in terms of the number of nodes it contains and represents the complexity of the tree. A tree of size 1 is equivalent to a simple rule which predicts the same outcome for all instances; a tree of size 3 contains one decision point with two branches which contain one leaf each; and so on.

For the target's colour and location, neither the pruned nor the unpruned trees improve on the baseline on either corpus. Looking at the trees produced shows that, not surprisingly, they are identical to the simple majority class rules of the baseline, which always include tg_colour and never include tg_location. With the exception of tg_colour in GRE3D3, these simple rules already achieve over 97% accuracy, which shows that people mostly used tg_colour regardless of its discriminatory power or visual salience and that they almost never used tg_location despite the fact that it was very useful in terms of discriminatory power, especially in the smaller GRE3D3 scenes.

The results for the always-include baseline for tg_colour demonstrate that the

	training corpus	tree size pruned	tree size unpruned
target colour	GRE3D3	1	1
	GRE3D7	1	1
target size	GRE3D3	3	3
	GRE3D7	11	13
target location	GRE3D3	1	1
	GRE3D7	1	1
relation	GRE3D3	5	9
	GRE3D7	1	1

Table 6.5: The sizes of the decision trees for target attributes. Size is measured in number of nodes in the tree.

participants used tg_colour in almost all trials in GRE3D7 (98%), but not in GRE3D3 (78%). This difference might be due to the fact that in order to describe the target without using tg_location or relation, i.e. using a inherent minimal description[4], tg_colour is necessary in 62.5% of the scenes in GRE3D7, but only in 40.0% of the GRE3D3 scenes. It might be the case that seeing more cases in which colour is necessary prompts people to use it more overall, even when they could describe an object without using colour. However, no decision trees can be constructed which capture more of the variation than the baseline rule; so, those participants who varied their use of tg_colour between trials were not guided in their decision by the cues from the scene which are encoded in my features.

The lack of use of tg_location in the GRE3D3 Corpus might be due to the fact that the most likely description of the target's location was similar to the relation between the target and the landmark. This might prompt people to use the full relation as in Example (6.4), instead of only a locational expression as in Example (6.5).

(6.4) the ball in front of the cube

(6.5) the ball in the front

In the GRE3D7 Corpus the location of an object was less clear cut and therefore harder to express. A simple locational expression, such as *in the centre* or *in the front*, which in GRE3D3 would suffice to fully identify an object, would only narrow the set of possible candidates down to two or three objects in GRE3D7.

Similarly as for the target's location, the classifier is also not able to find any rules that improve the prediction accuracy over that of the baseline rule for

[4]See Section 3.2.2 for a discussion of this term.

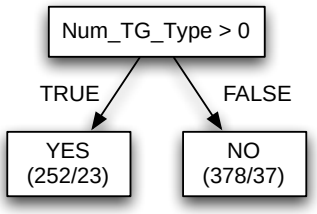

Figure 6.5: The tree characterising the use of tg_size **in GRE3D3.** Pruning had no effect on this tree.

relation in the GRE3D7 Corpus, which means that it never gets chosen. Because relation was used more often than tg_location in this corpus, the majority class rule for relation achieves lower results than that for tg_location. From the size of the tree for GRE3D3, we can see that it is more complex than the simple baseline rule. However, this tree does not perform significantly better than the baseline in characterising the use of relation. There only is a slight increase for the unpruned tree as well as for testing on the training set instead of using ten-fold cross-validation.

The most interesting results are achieved by the decision trees that characterise the use of the target's size. In both corpora, the decision tree performs vastly better than the majority class baseline. The baselines for this attribute are very low in both corpora, showing that there is no extreme preference for or against using tg_size in general as there is for tg_colour and tg_location. The high accuracy of the decision trees then shows both that the use of tg_size highly depends on the appearance of the scene, and that the features that I chose to characterise the scenes for the machine learner capture the aspects of the scenes' appearance that play a role in people's decision to use this attribute.

If we look at Table 6.5, it is interesting to note that the GRE3D3 trees for tg_size are less complex than those for relation, but nonetheless perform better. This indicates that the features available to the classifier in this experiment are much better suited to capturing the use of tg_size than that of relation in this corpus. The same is probably true for the GRE3D7 Corpus. However, the low use of relations overall in this corpus prevented the classifier from introducing any decisions into the tree for this attribute. Simply not using relation at all gives better results than the use of any combination of the scene-based features could. As we will see in Section 6.5.1, the decision to include a relation is much more dependent on the preferences of a particular participant than of the scene-dependent features.

Figure 6.5 shows the decision tree for tg_size in the GRE3D3 corpus (pruning

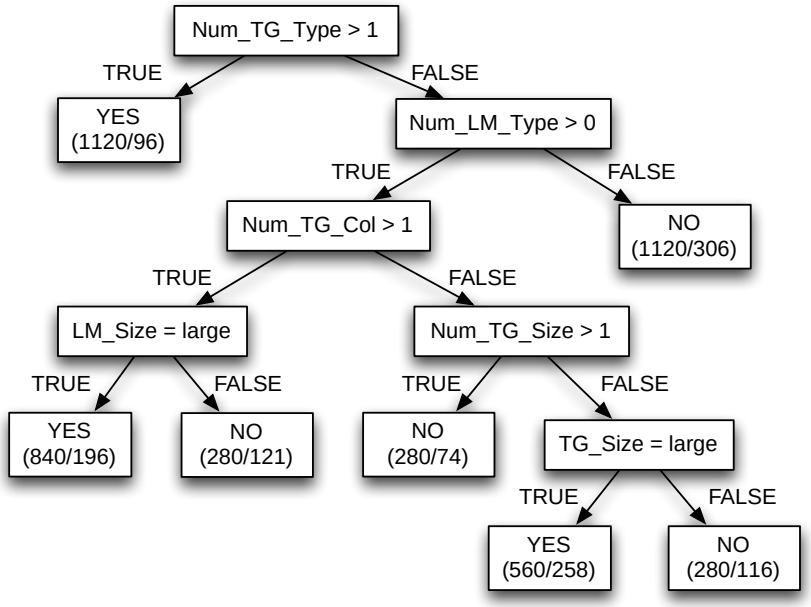

Figure 6.6: The unpruned tree characterising the use of tg_size **in GRE3D7.**

did not change the tree). Not surprisingly, it is exactly the same tree as the one for the prediction of the whole content pattern from Figure 6.3. The only difference between the two leaves of the pattern tree was that one of the predicted patterns contained tg_size and the other did not. The tree for predicting tg_size makes exactly the same distinction.

The unpruned tree characterising the use of tg_size in the GRE3D7 Corpus has two more nodes than the pruned one. It is shown in Figure 6.6. It too shows great similarity to the tree for predicting the full content pattern for the same corpus, with the first 4 splits being identical. Again, this is not highly surprising, as the only difference between the two patterns that were predicted in Section 6.2 was that one included tg_size and the other did not. The GRE3D3 tree for tg_size does not, however, confirm the hypothesis that people used size more in Scene 17 due to a curiosity in the stimulus design which made one large object placed further back look smaller. The content pattern tree used a split on the type of the relation between target and landmark to isolate that scene and assign to it the content pattern including tg_size (see Section 6.2). The tree in Figure 6.6 does not show such a split.

6.3.2 The Landmark's Attributes

As mentioned in Section 6.1.2, the descriptions used for these machine learning experiments only included attributes for the target object itself and the landmark object which was closest to the target. It is therefore possible to train decision trees to characterise not only the use of each of the target's attributes but also to characterise the use of the landmark's attributes. To do this, it is necessary to take into account the fact that the number of descriptions that contain a relation was overall fairly low in both corpora. This low number of relational descriptions of course results in even lower numbers of descriptions that contain a given landmark attribute, as only a proportion of the descriptions mentioning the landmark will also mention each of its attributes.

In earlier work using only the GRE3D3 corpus (Viethen and Dale, 2008; Dale and Viethen, 2009), we simply allowed the decision tree learner to use a feature called Relation_In_RE when it was charged with producing a decision tree for the inclusion of a landmark attribute. This feature encoded whether a description contained a relation to the landmark. Not surprisingly, all decision trees included this feature and only made further decisions about a landmark attribute after checking whether a relation was used.

Using this approach gives the decision trees a clear advantage over the baseline, as they can include knowledge about the description which the majority class baseline has no access to: the decision trees essentially cheat by taking into account information about the referring expression they are helping to construct. Simply because the majority of descriptions did not contain a relation at all, the baselines never included any of the landmark's attributes, while the decision trees were allowed to first hone in on the relational descriptions only and then make a decision.

As an alternative approach, I here present the accuracy results for both baseline and decision trees when trained on the relational descriptions only. These numbers are of course not directly comparable to the numbers for the target attributes, as they are based on a much smaller set of instances: 224 from GRE3D3 and 600 from GRE3D7, compared to the 630 and 4480 descriptions that the two corpora contain overall. However, they give a much clearer picture of the success in characterising the use of landmark attributes because they are not boosted by the large number of non-relational instances which both baseline and decision trees can easily predict correctly. This approach also allows for a fairer comparison between the results of the baseline and the decision trees.

Table 6.6 shows the accuracy that was achieved by the trees characterising the participants' use of the landmark's attributes in a description based on the scene's

	training corpus	test method	baseline	pruned tree	unpruned tree
landmark colour	GRE3D3	10 fold X	70.09%	67.41%	67.86%
	GRE3D3	training set	70.09%	70.09%	71.43%
	GRE3D7	10 fold X	86.83%	86.83%	86.83%
	GRE3D7	training set	86.83%	86.83%	86.83%
landmark size	GRE3D3	10 fold X	69.20%	**86.16%**	**86.16%**
	GRE3D3	training set	69.20%	**86.16%**	**86.16%**
	GRE3D7	10 fold X	54.50%	55.17%	56.17%
	GRE3D7	training set	54.50%	61.33%	62.00%
landmark location	GRE3D3	10 fold X	79.46%	**89.73%**	**89.73%**
	GRE3D3	training set	79.46%	**89.73%**	**89.73%**
	GRE3D7	10 fold X	98.33%	98.33%	98.33%
	GRE3D7	training set	98.33%	98.33%	98.33%

Table 6.6: Accuracy for the trees characterising the use of each landmark attribute trained only on instances containing a relation. As the results are percentages of the number of relational descriptions rather than all descriptions, they are not directly comparable to the results for the use of the target's attributes. (Bold values are statistically significantly different from the baseline at $p<0.02$ using the χ^2 test.)

characteristics. Table 6.7 lists the sizes of the trees.

The first observation to be made here is that the majority class baseline as well as the pruned and the unpruned trees perform on average worse in characterising the use of the landmark's attributes than those trained to predict the use of the target's attributes (c.f. Table 6.4). The lower baseline results simply mean that the participants used the landmark attributes less than the target attributes. While we have to keep in mind that these numbers are based on many fewer instances, the lower accuracies of the decision trees seem to indicate that the use of the landmark attributes is less dependent on the scene-based features I defined than the use of the target's attributes.

We see an improvement over the baseline for all decision trees except the ones for lm_colour and the GRE3D7 tree for lm_location. The landmark's location was used so rarely in GRE3D7 that simply never including it results in 98.33% accuracy, a baseline that is almost impossible to beat. The landmark's colour, on the other hand, was used so often that simply always including it achieves the best results. The scene-based features afford no help in characterising the cases that the baseline gets wrong. Even the relatively large unpruned tree predicting the use of lm_colour the GRE3D3 only performs very slightly better than the baseline, and only when tested on the full training set. It identifies TG_LM_Same_Size and TG_LM_Same_Col

	training corpus	tree size pruned	tree size unpruned
landmark colour	GRE3D3	1	9
	GRE3D7	1	1
landmark size	GRE3D3	3	3
	GRE3D7	11	23
landmark location	GRE3D3	3	3
	GRE3D7	1	1

Table 6.7: The sizes of the decision trees for landmark attributes based on relational descriptions only. Size is measured in number of nodes per tree.

as the main decision points for peoples' choice as to whether to use lm_colour, both features comparing the target's and landmarks attributes directly. Similarly as for the target's colour, performance for the landmark's colour is better on GRE3D7 than on GRE3D3, which shows that the participants' use of the landmark's colour follows similar principles as their use of the target's colour. In ten-fold cross-validation, even the pruned GRE3D3 tree for tg_colour performs worse than the baseline, although the tree trained on the whole set is identical to the majority class baseline rule. This must be due to the fact that at least one fold resulted in a different tree from this rule, which introduced additional error.

The only other tree for which pruning made any difference was the GRE3D7 tree for the landmark's size, which was pruned from 23 nodes to 11. The fact that, even when tested on the training data, the unpruned tree does not achieve much better results than the pruned tree means that even for the purpose of obtaining a clearer picture of which features best capture the variation in the use of landmark size, looking at the pruned tree suffices (see Figure 6.7). The first split in this tree tests for the type of relation between target and landmark object, showing that the majority of descriptions with a horizontal relation use the landmark's size, while the case for the descriptions with vertical relations is more complicated: the next two splits test for the number of objects sharing size with the target object. Only at the fourth level of the tree does a feature appear that has to do explicitly with the landmark's size: Num_LM_Size, the number of objects sharing their size with the landmark.

The GRE3D3 decision tree for the landmark's size, shown in Figure 6.8, is comparatively simple (pruning had no effect for this tree). It only checks whether the target and landmark have the same colour. If this is the case, lm_size is included. This makes sense intuitively, as lm_colour in this case would not distinguish the landmark from the closest object, the target, which might prompt people to

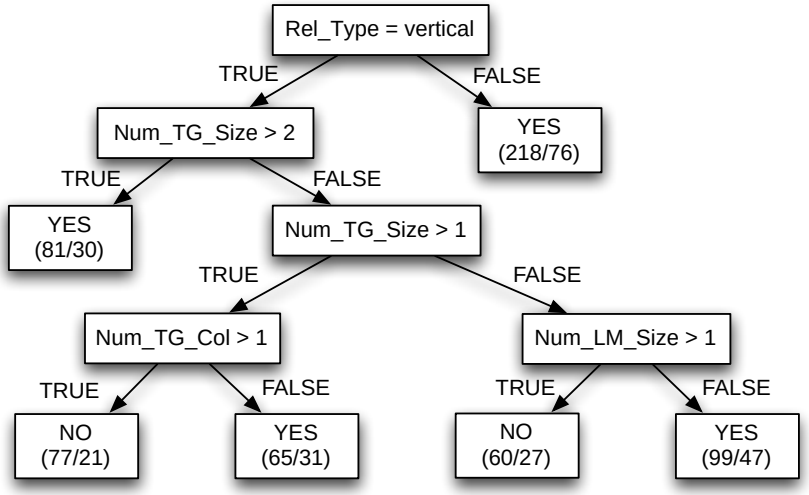

Figure 6.7: The pruned decision tree characterising the use of lm_size in GRE3D7.

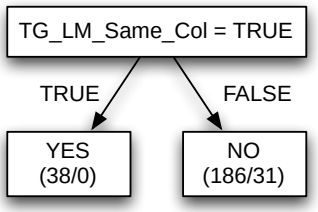

Figure 6.8: The decision tree characterising the use of lm_size in GRE3D3. Pruning had no effect on this tree.

include lm_size instead of or as well as lm_colour.

Interestingly, the GRE3D3 tree for lm_location achieves quite an improvement over the baseline compared to its counterpart for the target's location, which simply mimicked the baseline rule. It is shown in Figure 6.9. It checks whether the target shares its colour with any objects. If it does, the tree predicts that people do not use the landmark's location. If the target has a unique colour (which means that the scene was based on design Schema A from Section 5.2.1) the landmark's location is included.

A possible explanation for the success of this strategy is that the landmark and distractor object look identical in scenes based on Schema A, which means that the only way to distinguish the landmark from the distractor is by using its location left or right in the picture. It is interesting that the participants found it

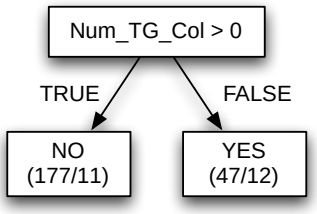

Figure 6.9: The pruned decision tree characterising the use of lm_location in GRE3D3.

necessary to do so, considering that both the target and the landmark are already uniquely identified by the spatial relation between them.

This lends support to recursive approaches to relational reference which start describing a freshly introduced landmark 'with a blank slate' and do not take into account the partial description of the target referent already constructed. Here, the new object is not assumed to be already partially distinguished by its relation to the target object and the attributes included for it, but rather has to be distinguished in its own right. This results in descriptions such as

(6.6) the rabbit in the hat on top of the table

in a situation where there is another rabbit that is not inside a hat and another hat without a rabbit inside it and also not on a table.

6.4 Cross-Corpus Testing

In a separate experiment, I tested the decision trees trained on each corpus on the respective other corpus to assess how different the two corpora really are from each other. Table 6.8 compares the accuracies of testing the majority class baseline rules and the pruned trees using cross-corpus testing to the ten-fold cross-validation results reported in the previous two sections. It also shows the sizes of the pruned trees again.

For cross-corpus testing some of the same patterns apply as for ten-fold cross-validation. Firstly, pruning makes no difference to the performance —mostly because hardly any pruning is performed— which is why the table omits the results for the unpruned trees. Secondly, only the trees predicting the use of the whole content pattern and of the target's and landmark's size achieve an improvement over the baseline. This improvement is particularly pronounced for the GRE3D7 tree that characterises the use of tg_size when tested on the GRE3D3 data. The

learned item	training corpus	test method	baseline	pruned tree accuracy	pruned tree size
pattern	GRE3D3	cross-corpus	36.70%	**47.88%**	3
		10-fold X	27.30%	**46.51%**	
	GRE3D7	cross-corpus	22.70%	**36.98%**	15
		10-fold X	47.88%	**64.71%**	
target colour	GRE3D3	cross-corpus	98.73%	98.73%	1
		10 fold X	77.94%	77.94%	
	GRE3D7	cross-corpus	77.94%	77.94%	1
		10 fold X	98.72%	98.72%	
target size	GRE3D3	cross-corpus	42.25%	**57.75%**	3
		10 fold X	57.78%	**90.48%**	
	GRE3D7	cross-corpus	42.22%	**90.48%**	11
		10 fold X	57.75%	**73.95%**	
target location	GRE3D3	cross-corpus	98.19%	98.19%	1
		10 fold X	97.62%	97.62%	
	GRE3D7	cross-corpus	97.62%	97.62%	1
		10 fold X	98.19%	98.19%	
relation	GRE3D3	cross-corpus	86.61%	86.61%	5
		10 fold X	64.44%	64.44%	
	GRE3D7	cross-corpus	64.44%	64.44%	1
		10 fold X	86.61%	86.61%	
landmark colour	GRE3D3	cross-corpus	86.83%	86.83%	1
		10 fold X	70.09%	67.41%	
	GRE3D7	cross-corpus	70.09%	70.09%	1
		10 fold X	86.83%	86.83%	
landmark size	GRE3D3	cross-corpus	45.50%	47.00%	3
		10 fold X	69.20%	**86.16%**	
	GRE3D7	cross-corpus	30.80%	**57.59%**	11
		10 fold X	54.50%	55.17%	
landmark location	GRE3D3	cross-corpus	98.33%	98.33%	3
		10 fold X	79.46%	**89.73%**	
	GRE3D7	cross-corpus	79.46%	79.46%	1
		10 fold X	98.33%	98.33%	

Table 6.8: Accuracy for the trees characterising the use of whole content patterns and each property in cross-corpus testing. Bold values are statistically significantly different from the baseline at $p<0.03$ using the χ^2 test.

GRE3D7 baseline rule only predicts 42.22% of GRE3D3 cases correctly, however the decision tree is, at 90.75% accuracy, just as good at characterising the use of tg_size in GRE3D3 as the tree trained on GRE3D3 itself.

The main observation from this table is that cross-corpus testing achieves surprisingly high accuracy scores, which are in many cases even better than the scores for ten-fold cross-validation on the training set. This means that in these cases the trees are worse at capturing the variation found in the corpus they were trained on than that in the other corpus. This is especially the case for trees trained on GRE3D3. Only the GRE3D3 trees for tg_size and lm_size perform better on GRE3D3 itself than on GRE3D7. Interestingly, the trees characterising the use of tg_size and lm_size are the only trees trained on GRE3D7 that perform better in the cross-corpus testing than in ten-fold cross-validation.

The fact that the GRE3D7 trees for size capture the GRE3D3 data better than the GRE3D7 data itself and that the GRE3D7 tg_size tree does just as well on the GRE3D3 data as the GRE3D3 tg_size tree again indicate the interesting variation we have seen before for size. The baseline results for both the target's and the landmark's size are below 50% in cross-corpus testing for both corpora. This shows that the trend for the use of size is exactly the opposite in the two corpora. From the quantitative analysis in Chapters 5 we know that both tg_size and lm_size are contained in a majority of the descriptions in GRE3D7, while for GRE3D3 the opposite is the case. The better cross-corpus results and larger sizes of the decision trees trained on GRE3D7 suggest that the usage patterns for size are more complex in this corpus and subsume the usage patterns in GRE3D3. Of course, this could be due to the fact that GRE3D3 is a much smaller corpus than GRED7; however, if this was the case, one might expect to see the same result patterns for the other properties and also for the complete content patterns, which is not the case.

For all other properties, the general tendency of GRE3D3 trees to perform better on GRE3D7 than on GRE3D3, while GRE3D7 trees tend to perform better on GRE3D7 itself, reflects the high variability of GRE3D3 discussed before. It seems that from GRE3D3 we can learn rules that subsume the variability of GRE3D7, while the opposite is generally not the case. The fact that the performance of most GRE3D3 trees increase when tested on GRE3D7 rather than just staying the same, leads to the conclusion that either GRE3D3 is too hard to predict or that my features capture the variation of GRE3D7 better than that found in GRE3D3.

training corpus	test method	−Participant_ID pruned Acc	−Participant_ID pruned size	+Participant_ID pruned Acc	+Participant_ID pruned size	+Participant_ID unpruned Acc	+Participant_ID unpruned size
GRE3D3	10 fold X	46.51%	3	54.60%	415	**57.46%**	573
	training set	46.51%		**91.27%**		**98.10%**	
GRE3D7	10 fold X	64.71%	15	67.01%	1023	63.71%	2798
	training set	64.93%		**82.59%**		**93.77%**	

Table 6.9: Accuracy for learning the use of whole content patterns based on scene and participant information. Bold values are statistically significantly different to the participant-insensitive trees at $p<0.03$ using the χ^2 test.

6.5 Speaker-Dependent Variation

In this section, I investigate how much of the variation in the corpora is due a speaker's preferences. I re-run the learning experiments from Sections 6.2 and 6.3 with Participant_ID as an additional feature for the machine learner to choose from for the trees. This allows a direct comparison of the scores the trees achieve with and without this feature. The difference between these scores gives a clear indication of how much variation is due to participants' preferences. The results of this experiment are presented in Section 6.5.1

I am also interested in finding out how many people do the same thing for each property (Section 6.5.2). This is much easier to compute if I have individual trees, one per participant and property, rather than distilling the behaviour of each participant from one large tree per property. Using these speaker-specific trees I motivate the notion of a **speaker profile**, consisting of the complete set of trees for this speaker.

6.5.1 Speaker as a Prediction Feature

Again, I first trained trees for the prediction of the whole content pattern. The accuracy results for these trees are summarised in Table 6.9. Three main observations are to be made: Firstly, when the decision tree learner has the option to include Participant_ID as a feature, it does, and this results in very large trees. Secondly, including participant as a feature helps significantly to characterise the use of full content patterns in the two corpora, although only the unpruned tree trained on GRE3D3 achieves a significant improvement in ten-fold cross-validation. And thirdly, the trees using Participant_ID as a feature perform vastly better on the complete training set than when tested using ten-fold cross-validation.

The last observation is probably due to a data sparseness problem that leads to overfitting to the data: GRE3D3 contains only 10 descriptions from each par-

ticipant, GRE3D7 has 16 descriptions per participant, and in both data sets each person gave only one description for each scene they saw. A tree trained on only a part of the limited data from one participant is unlikely to capture a usage pattern for this participant which also fully characterises this participant's remaining data, which will be used for testing. However, if I were able to collect more samples for similar stimulus scenes from the same person, it is likely that the tree trained on the full data set would be able to predict this new data more accurately than the ten-fold cross-validation results.

Overall, this shows that a lot of the variation in content pattern use is due to personal preferences that vary from speaker to speaker. It also shows that the features I used to characterise the scenes are successful at capturing the within-participant variation in the data. Finally, it shows that the personal preferences of my participants vary too much within the set of descriptions each participant gave to allow generalisation to unseen but similar situations.

Again, I am interested to see how well trees only charged with deciding about the inclusion of a single attribute perform. Table 6.10 compares the accuracy for predicting the use of the target's attributes of pruned and unpruned trees which were built taking into account the Participant_ID to the respective pruned trees without participant information.

Looking at the sizes of the trees, we see that pruning reduces the trees for tg_location and on GRE3D7 also those for tg_colour and relation back down to the baseline rules, which never include location or relation and always include colour. Shifting our attention to the performance of these rules in ten-fold cross-validation, we see that they achieve the same accuracy or slightly better than the large unpruned trees which are participant-sensitive. In fact, this is true for all attributes on both corpora.

As for predicting the whole content pattern, very high performance on the training set in most categories reveals that the unpruned trees in particular are very good models for the data. Considering that the unpruned trees contain much more personalised information about each participant, this again underlines the fact that much of the variation found in the two corpora is due to speaker-specific preferences.

The results for the trees predicting the inclusion of the landmark's attributes show the same trends: testing on the whole training set, the unpruned trees do very well, while the results for ten-fold cross-validation are lower.

	training corpus	test method	−Participant_ID pruned Acc	−Participant_ID pruned size	+Participant_ID pruned Acc	+Participant_ID pruned size	+Participant_ID unpruned Acc	+Participant_ID unpruned size
target colour	GRE3D3	10 fold X	77.94%	1	**90.16%**	100	**89.21%**	201
		train set	77.94%		**96.98%**		**98.73%**	
	GRE3D7	10 fold X	98.73%	1	98.73%	1	98.04%	429
		train set	98.73%		98.73%		99.71%	
target size	GRE3D3	10 fold X	90.48%	3	90.48%	3	90.48%	219
		train set	90.48%		90.48%		**99.05%**	
	GRE3D7	10 fold X	73.95%	11	77.68%	805	76.21%	2003
		train set	73.95%		**89.06%**		**94.82%**	
target location	GRE3D3	10 fold X	97.62%	1	97.62%	1	97.14%	102
		train set	97.62%		97.62%		**99.37%**	
	GRE3D7	10 fold X	98.19%	1	98.19%	1	97.08%	1021
		train set	98.19%		98.19%		99.53%	
relation	GRE3D3	10 fold X	64.44%	5	**82.06%**	134	**82.06%**	200
		train set	65.87%		**94.76%**		**97.78%**	
	GRE3D7	10 fold X	86.61%	1	86.61%	1	86.52%	1343
		train set	86.61%		86.61%		**98.35%**	
landmark colour	GRE3D3	10foldX	67.41%	1	**84.38%**	96	**84.82%**	116
		training set	70.09%		**97.32%**		**100%**	
	GRE3D7	10foldX	86.83%	1	86.83%	1	86.83%	411
		training set	86.83%		86.83%		**100%**	
landmark size	GRE3D3	10foldX	86.16%	3	86.16%	3	89.29%	157
		training set	86.16%		86.16%		**99.11%**	
	GRE3D7	10foldX	55.17%	11	**70.33%**	359	**67.33%**	485
		training set	61.33%		**93.00%**		**97.5%**	
landmark location	GRE3D3	10foldX	89.73%	3	89.73%	3	88.39%	151
		training set	89.73%		89.73%		**100%**	
	GRE3D7	10foldX	98.33%	1	98.33%	1	97.17%	307
		training set	98.33%		98.33%		99.67%	

Table 6.10: Accuracy for the trees characterising the use of each attribute based on scene and participant information. Bold values are statistically significantly different to the participant-insensitive trees at $p<0.02$ using the χ^2 test.

6.5.2 Training Speaker-Specific Trees

Of course, it would be possible to read the strategies characterising each participant's behaviour off the trees described in the last section. However, these trees are often extremely large, as they split at least once on the participant feature. To be able to compare the behaviours that were learned for different participants for each attribute more easily, I instead trained speaker-specific trees. To reduce the number of trees to train, and thereby the number of comparisons between trees, I only used pruned trees in these experiments. Each tree represents a strategy that the corresponding speaker seems to follow when deciding whether to include

the relevant attribute or not. To train these trees, the machine learner is given the same features to choose from and has to make the same predictions as in the last section, but it only works on the data of one participant at a time. In this way, individual strategies for a particular participant for a particular attribute are produced, which can be compared easily to the strategies for other participants. It is then straightforward to compute how many participants do the same thing overall and for a given attribute.

It also makes it possible to collect for each participant a complete set of trees, one per attribute. I will call such a collection the **speaker profile** of a participant. A speaker profile can act as a replacement for the tree learned for the prediction of the content patterns a participant used as it has the same function: to predict the full content of a description given the features of the scene and the person who produced it.

Speaker profiles have two main advantages over trees predicting content patterns: First, their modularity allows us to find cross-participant commonalities in attribute-specific behaviour that is hidden in trees which predict complete content patterns. For example, participant A might share his strategy for including colour with one set of other participants and use the same strategy for including relations as another set of participants, but this participant A is the only one that combines these two strategies, leading to a unique overall behaviour. A's tree for the prediction of complete content patterns will be unique from all others and it would be very hard to discern from it that he actually shares some attribute-specific behaviours with other participants.

Second, attribute-specific trees for individual speakers facilitate a more fine-grained definition of error. If a tree makes a certain percentage of incorrect predictions, all we know is in how many cases it did not choose exactly the same content as the descriptions in the corpus; but we do not know is whether it only got one attribute wrong in each predicted referring expression or maybe all of them. Looking at the combined error of all attribute-specific trees in a speaker profile affords a much stricter definition of error: each time the profile predicts a single attribute incorrectly, the error count increases by one. So, the definition of error is shifted from the level of complete descriptions to the level of individual attributes. The attribute-specific speaker-sensitive trees conveniently come with an error count for this particular attribute. This makes it easy to read off how often a complete speaker profile gets a single attribute wrong, and to calculate how often it would get any attribute wrong over the course of the whole set of descriptions from the corresponding participant.

I first look at the number of different trees that were built for each attribute

	absolute counts		% out of maximum possible number	
	GRE3D3	GRE3D7	GRE3D3	GRE3D7
target colour	7	7	11.11%	2.50%
target size	9	50	14.29%	17.86%
target location	2	10	3.17%	3.57%
relation	17	33	26.98%	11.79%
landmark colour	12	11	19.05%	3.93%
landmark size	8	13	12.70%	4.64%
landmark location	5	1	7.94%	0.36%
full pattern	42	102	66.67%	36.43%

Table 6.11: The number of speaker-specific trees per attribute. The left half shows the absolute counts and the right half shows the numbers as a percentage of the maximum possible number of trees, which is equal to the number of participants (63 for GRE3D3 and 280 for GRE3D7).

to see how often participants seem to be following the same strategies for these attributes. Following this, I move on to an examination of the error rate of the participant-specific trees and the speaker profiles.

Table 6.11 lists for each attribute how many different speaker-specific trees were trained. For completeness, the table also includes the counts of trees for the complete content pattern. Comparing the two columns shows that in GRE3D3 most between-speaker variation occurs in the use of relation and lm_colour. In GRE3D7, the machine learner produces by far the most trees for tg_size and relation. So, the two corpora share a spike in between-speaker variation for the use of relation, but differ in their second spike. These spikes are congruent with the largest increases in performance when allowing Particpant_ID as a feature that we saw in Table 6.10: the use of these attributes is particularly dependent on the preferences of the speaker.

It is not surprising to see that for almost all attributes there are more different strategies in the GRE3D7 corpus, as almost four and a half times more participants contributed to this corpus than to GRE3D3. However, in the right half of Table 6.11, the number of different strategies are shown as a percentage of the maximum possible number of strategies. This is the same as the number of participants, as the maximum number of trees would result from each participant following a different strategy. Normalised by the number of participants, it appears yet again that GRE3D7 is less varied than GRE3D3, as the percentages for GRE3D7 are almost all lower than those for GRE3D3.

One interesting observation that starts to emerge from this data is that, despite

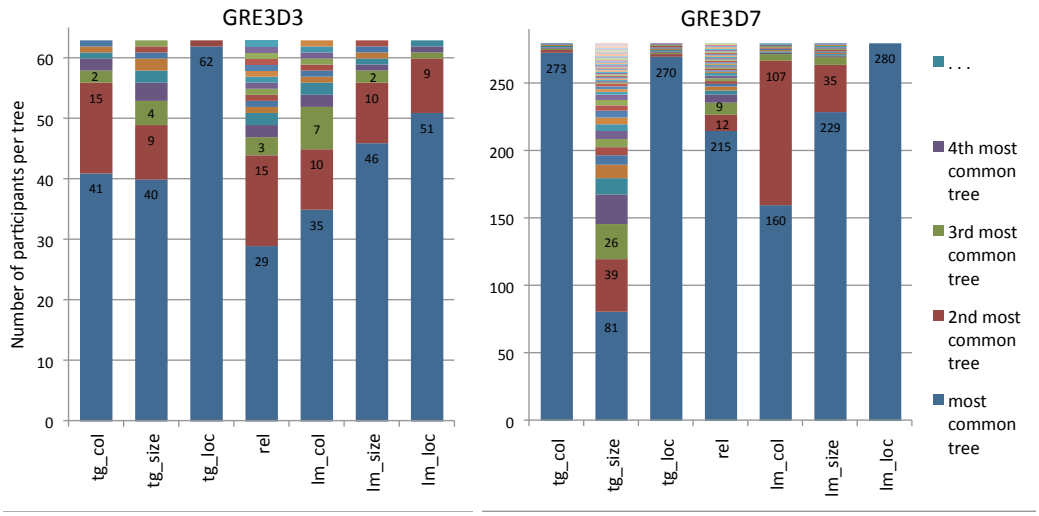

Figure 6.10: Number of participants sharing each attribute-specific decision tree. The more different colour slices in a column, the more different trees exist for this attribute. The thinner a colour slice, the fewer participants shared the corresponding tree.

the fact that speaker preferences clearly play a major role in content determination for referring expressions, it is by no means the case that there are no commonalities between the contributors to the corpora. If every speaker's behaviour was different, there would be as many trees as there are participants; and while the number of different trees (i.e. different behaviours) at the level of full content patterns is fairly high, in particular in the GRE3D3 corpus, breaking down the content patterns into the individual attributes shows that at this level the participants' reference behaviour varies much less. Of course, this is not a great surprise, considering the increase in performance that we saw throughout this chapter every time we moved from predicting whole content patterns to predicting individual attributes.

The graphs in Figure 6.10 visualise how many participants actually shared strategies for each attribute. Each column shows for a given attribute how many participants share each strategy that was learned for this attribute. The most common strategy is represented by the longest bar at the bottom, the second most common strategy by the second longest bar stacked on top of the bottom bar, and so on. The actual counts are shown inside the bars for the most common strategies. The graphs clearly show that for all attributes in both corpora, the majority of participants used a strategies which is shared by a substantial number of participants. In other words, while there is substantial between-speaker variation, the majority of participants do not follow a behaviour pattern that is unique to them. These graphs confirm that the highest between-speaker variation in GRE3D3 occurs for

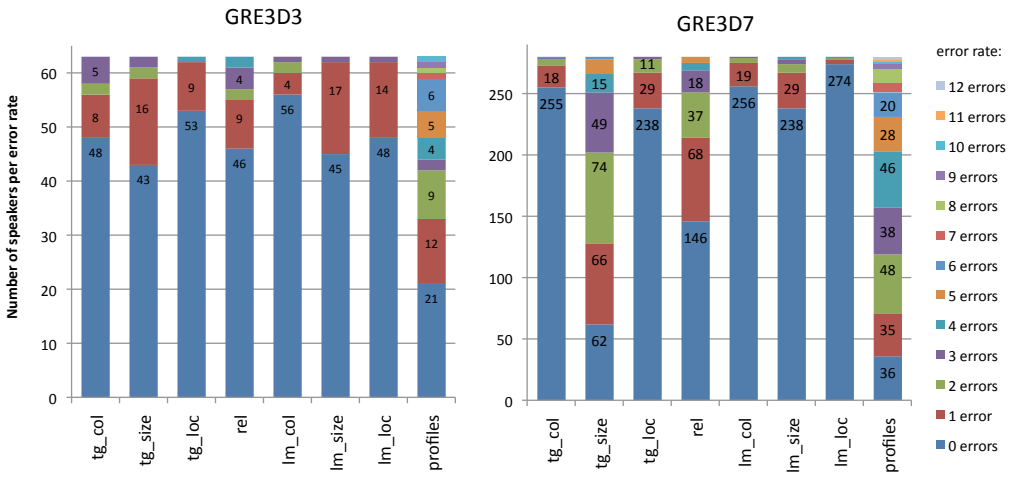

Figure 6.11: Number of participants whose trees had a given error rate.

relation, while for GRE3D7 tg_size shows even higher between-speaker variation. The actual counts for all trees are shown in Table C.1 in Appendix C.

We already know which attributes are harder to predict than others from the accuracy that the participant-specific trees achieved (see Table 6.10). However, it would also be interesting to know how reliable the individual strategies for each person are. It might be that the behaviour of some participants is harder to predict on the basis of my scene-dependent features than that of others. To investigate this question, I introduce the notion of the **error rate** of an attribute-specific tree of a given participant. For example, if the tree which was learned to predict the inclusion behaviour of a given participant for tg_size makes three mistakes over all referring expressions produced by this participant, then the error rate for tg_size for this participant is three. The maximum error rate for a tree in GRE3D3 is ten, as each participant contributed ten descriptions, while in GRE3D7 the maximum error rate per tree is 16. This also makes it possible to calculate the number of participants whose trees had a given error rate.

The graphs in Figure 6.11 show for a given error rate and a given prediction class the number of participants whose trees or profiles had this error rate. (Table C.2 in Appendix C shows the exact counts for all error rates.) The (blue) bottom bars represent the number of participants whose decision trees made no prediction error, the (red) bars on top of that represent the number of participants for which one error was made, and so on. So, for example in GRE3D3, tg_colour was predicted by a tree making zero errors for 48 participants. We can see that

most attributes can be predicted without error for the majority of participants, with the exception of tg_size in GRE3D7. This shows that there is a majority of participants whose behaviour in terms of including most individual attributes is easy to predict from the data. The errors made by the trees must therefore stem from a minority of hard-to-predict participants. A closer inspection also reveals that, in both corpora, the speaker profiles predict the behaviour of less than 45% of all speakers with an error rate of four or more (19 out of 63 speakers in GRE3D3 (30%) and 123 out of 280 speakers in GRE3D7 (44%)). So, in both corpora less than 45% of the speaker profiles are responsible for more than 70% (75.5% in GRE3D3 and 73.8% in GRE3D7) of the error made by all speaker profiles overall.

6.6 Discussion

In this chapter, I have presented a series of machine learning experiments in which I trained decision trees aimed at characterising what content people choose to include in referring expressions when describing simple objects in small 3D scenes. I trained decision trees to predict the full content pattern of a given description and also to predict for each individual attribute whether it should be included or not. First, the decision trees were only given information about the scenes and the objects contained in them. In a second step, I included the participant ID as a feature. In the following I describe the conclusions that I draw from these experiments, followed by a discussion of the consequences that the experiments presented in this chapter have for the development of REG algorithms.

6.6.1 Conclusions

Complete descriptions are hard to predict, but more commonality exists between speakers at the level of individual attributes: The results for the trees characterising the use of full content patterns show that predicting the exact content of a referring expression based solely on the visual appearance of the objects in a stimulus scene is very hard. The experiments characterising the use of individual attributes showed that for the target's colour and location, as well as the relation between the target and the landmark, no feature-based decision tree could be constructed that would significantly outperform a simple majority class baseline. The use of the target's size is much more variable and highly dependent on the visual appearance of the scene as characterised by the features used in the decision trees.

For the landmark's attributes I found a similar pattern: the biggest improvements were made by the trees characterising the use of the landmark's size. The

decision trees for the landmark's location trained on the GRE3D3 data suggest that people felt the need to distinguish the landmark from the third object in the GRE3D3 scenes, despite the fact that the relation between target and landmark already identified the landmark exhaustively. This lends support to approaches to the generation of referring expressions which do not take into account the already included content of the description under construction when deciding on the inclusion of landmark attributes.

Cross-corpus testing confirms that GRE3D7 is less variable than GRE3D3: Testing the trees trained on each corpus on the other corpus brought additional confirmation of the higher variability of the smaller GRE3D3 corpus compared to GRE3D7, which was already evidenced in the previous chapter. With the exception of the trees characterising the use of size, trees trained on GRE3D3 achieved higher Accuracy when tested on GRE3D7 than when tested on GRE3D3 itself, while those trained on GRE3D7 also performed better on GRE3D7. The usage patterns in GRE3D3 therefore seem to subsume the usage patterns of GRE3D7, while the opposite is not the case.

Using Participant_ID **as a feature produces trees that characterise the data set well, but do not necessarily generalise to new data:** Once information about the participant who produced a given referring expression was taken into account, it was possible to train decision trees on the scene-based features which characterise the use of full content patterns with much higher Accuracy. When tested on the full training data, participant-specific trees achieved more than 90% Accuracy on both corpora for the prediction of which content pattern to use. This suggests that participant-specific variation is the most influential factor. It is likely that speaker identity was also the best predictor of semantic content in Jordan and Walker's (2005) machine learning experiment, although they only tried it as one of a set of several discourse-independent features (they called them 'inherent' features), also including the attribute values of the target referent and the utterance number within the dialogue. Using only this set of features they achieved better results than with any discourse-theoretical features.

Speaker profiles offer a method to capitalise on the commonalities between speakers at attribute level in a new way of modelling REG**:** I found that attribute-specific trees which take into account participant information achieve higher Accuracy scores than the participant-specific trees characterising the use of complete content patterns, even when tested under ten-fold

cross-validation. An analysis of the attribute-specific trees for each individual participant showed that most people share each of their attribute-specific strategies with a large number of other participants and only very few people 'do their own thing'. This means that, while at the level of entire referring expressions a lot of cross-speaker variation makes it difficult to characterise human reference behaviour, a great deal of commonality exists in the strategies used for the inclusion of individual attributes.

A useful way of modelling people's strategies for building referring expressions is therefore to collect the attribute-specific trees of each participant into a speaker profile which then predicts complete referring expressions. These speaker profiles have two advantages over trees that predict entire content patterns at once: (1) they make it easier to compute a more fine-grained notion of error, and (2) they retain access to the individual attribute-specific strategies and the commonalities between speakers at that level.

Thinking about human referring behaviour in terms of speaker profiles amounts to a more bottom-up way of processing than what happens in traditional REG algorithms: rather than controlling the combination of individual attributes in serial dependency, where each decision about which attribute to include is impacted directly by the attributes that have been included already, the focus is on the individual attributes themselves. The picture that emerges is one where we can think of an individual speaker's approach to reference as consisting of a collection of attribute-specific strategies. The inclusion of each attribute is considered by an independent model that allows us to explore what it is that makes an attribute appropriate for inclusion in a developing referring expression in a given situation. The individual strategies are shared between speakers to a much greater degree than combinations of strategies, the speaker profiles. A speaker may even use different strategies depending on features orthogonal to the referential context (such as who they are talking to, or the mission-critical nature of getting it right first time).

6.6.2 Implications for Algorithm Development

In order to discuss the implications that the findings of this chapter might have for the development of REG algorithms, we have to distinguish between the different purposes that these algorithms can have. As discussed in Section 2.5, REG algorithms can have two very different aims. The first of these aims is for the algorithm to be able to serve in a real-world NLG application, such as a direction giving system in the virtual world of Stoia et al. (2006) or the robot of Kelleher and Kruijff (2006). These application-oriented algorithms necessarily place most

importance on the needs of a user. The second possible aim for a REG algorithm is to attempt to replicate the reference behaviour of speakers, in order to shed light on the processes that are involved in the production of referring expressions in natural language.

The focus of this thesis is on the second of these aims, so I begin by considering the consequences that the experiments described in this chapter might have for the development of algorithms that have the aim of replicating human reference behaviour. The easiest way to derive an actual algorithm from a model learned from data, such as the decision trees I trained in this chapter, is to simply use the decision trees themselves as the algorithm. The trees that predict the full content pattern could be used as is; their output is equivalent to that of traditional REG algorithms. As outlined in Section 6.5.2, the trees predicting whether each individual attribute should be used would have to be re-combined into what I call speaker profiles to generate full referring expressions.

One respect in which these decision trees are different from traditional REG algorithms, such as the IA or the graph-based algorithm discussed in Chapter 2, is the required representation of the underlying knowledge about the reference domain. Traditional REG algorithms rely on a knowledge base that simply lists for each object in the domain which value it takes for each of its attributes, and possibly in which relation, spatial or otherwise, the objects stand to each other. The algorithm then computes the differences between different objects and based on that makes its decisions as to which attributes should be included in a referring expression. For the machine learning experiments reported in this chapter I had to convert the direct information about the objects in a scene into machine learning features. Some of these features simply replicate the same direct object properties, but the more interesting features capture comparative properties, such as the number of objects that share a particular property with the target or whether target and landmark are of the same type.

This process shifts some of the burden from algorithm development to the development of the input representation for the machine learner, which will then build an algorithm automatically. While this might seem like an arbitrary choice between putting effort into designing an algorithm and putting effort into designing the input representation, doing the latter has one clear advantage in its flexibility. While the input for traditional algorithms was usually limited to the visual characteristics of the referential scenario, it is straightforward to include all manner of information in a set of machine learning features, such as, in this case, an identifier for the speaker whose reference behaviour is to be mimicked or, in the

case of subsequent reference, information about the discourse situation (Jordan and Walker, 2005). Including this type of information in the decision making process of traditional REG algorithms would require additional machinery, and it is not clear how it would best be integrated with the existing algorithms. A second advantage lies in the fact that the performance of the resulting system depends entirely on the input representation (i.e., the chosen feature set), rather than an intricate interplay between the algorithmic mechanism and the input representation.

If, instead, we take the perspective of application-oriented algorithm design, the main implication from the experiments in this chapter is that attempts to generate natural-sounding referring expressions by looking at human data need to account for the fact that people do different things. It seems unlikely that naturalness in the referring expressions produced by an application employed in a real life-situation can best be achieved by mimicking some sort of 'average' behaviour taking in characteristics of a diverse range of speakers.

Using the approach I took in Section 6.5, it is possible to extract the reference behaviour of individual speakers. Once these speaker-specific behaviours are characterised by decision trees, or other machine-learned models, it is possible to use them to let an application replicate one specific speaker, and thereby give it a 'personality'. Alternatively, one might want to choose, for example, the most common behaviour exhibited by individual speakers.

Of course, in an application it might be the case that the referring expressions to be produced need to meet certain hard criteria other than being human-like. One such criterion might be that each description needs to be fully distinguishing. However, when using a speaker profile combined from attribute-specific trees it cannot be guaranteed that the resulting referring expressions meet this criterion. It would therefore be necessary to build some machinery around the chosen speaker profile to ensure that mission-critical criteria are met.

One issue plaguing REG research, which the work I have presented in this chapter does not explicitly address, is the domain-specificity of algorithms. Unfortunately, REG strategies, such as preference orders for the IA or cost functions for the graph-based framework, cannot easily be ported to new, different, domains, and I did not aspire to solve this problem in this thesis. Decision trees, just as any other machine-learned models trained on actual data, are specific to the domain in which the data was collected and, as we have seen, to the participants from whom it was collected. Nonetheless, especially by using pruned decision trees one might hope that learned models might carry over at least to similar domains. However,

automatically learning models on one domain and then trying them out on another domain (as I did in Section 6.4, if one wants to call the GRE3D3 scenes a different domain from the GRE3D7 scenes) represents a low cost way of identifying whether two domains are similar, even if ultimately rule-based systems are to be used in an application. The decision trees resulting from two different domains can also serve as a powerful analytical tool by giving insights into the ways in which they are different, as only factors that impact the content of referring expressions produced by humans in each of them will be chosen as decision points in the trees, with the more important factors appearing closer to the root.

Chapter 7

Conclusions

7.1 Summary and Discussion

The research presented in this thesis was based on two main premises: that research in the computational generation of referring expressions should strive to achieve system output that is as human-like as possible; and that, to this end, we should endeavour to model human referring behaviour as it can be observed in corpora. Adopting these premises serves two purposes: Firstly, it improves the adequacy of the output of REG algorithms for object identification by mimicking the human ability to produce adequate references; and secondly, studying corpora of human-produced data and developing algorithms that can replicate this data might bring us closer to an understanding of what it is that humans do when they refer.

As I argued in the introduction to this thesis, the classic REG algorithms and most of their descendants were neither based on nor assessed against human-produced data. They were based on a rather minimalist view of what it takes for a referring expression to be optimal by concentrating on computational efficiency and short descriptions as their main concerns. A small number of existing approaches were based on observations about general human reference behaviour garnered from psycholinguistic experiments, but again they were not evaluated against human data.

The algorithms that were submitted to the REG evaluation challenges between 2007 and 2009 were, of course, tested on the TUNA Corpus, and some of them also took into account patterns found in the development set. Unfortunately, as I pointed out in Section 4.4.2, there are a number of concerns around the question of whether the TUNA Corpus, and the way the systems' output was compared to it in the challenges, are ideal for an assessment of the descriptive adequacy of REG systems. Nonetheless, it would be interesting to see the data-driven algorithms

from the challenges described in more detail and to evaluate them on a larger data set containing more than one referring expression for each stimulus item.

This thesis set out to tackle three main areas in which corpora can be used to further the aim of human-likeness in research on referring expression generation: corpus-based evaluation, corpus collection and analysis, and statistical modelling of corpus data. It started off with an appraisal of the state of the art in research in the generation of distinguishing descriptions, focussing in particular on work that has a bearing on the research presented in the later chapters, namely work dealing with spatial relations and work that has made use of corpora. This was followed by an examination of a number of methodological choices that have to be made when working with corpora in REG. Here, I explored the different options on offer for corpus collection exercises, which are centred around the balance that needs to be struck between controlling the experimental settings as much as necessary and keeping the settings as natural as possible. I discussed a number of concepts that are of import for analyses of REG corpora, such as the different nature of object properties, and the notions of minimality and over-specification of referring expressions. Finally, I analysed different ways in which a system's output can be compared to corpus data, under the premise that the aim of the comparison is to assess whether the system might be an descriptively adequate model of human reference behaviour.

The following three chapters described the research undertaken to address each of the three areas where corpora can be employed in REG: evaluation of human-likeness, corpus collection and analysis, and modelling of corpus data.

7.1.1 Corpus-Based Evaluation

The first of the three main content chapters presented an evaluation experiment in which three of the classical algorithms, Dale's (1989) Greedy Algorithm (GREEDY), Dale and Haddock's (1991b) Relational Algorithm (RA) and Dale and Reiter's (1995) Incremental Algorithm (IA), were put to the test regarding their ability to replicate the referring expressions found in a relatively small corpus of referring expressions in a grid-like visual domain of drawers in filing cabinets. The analysis of this experiment had two major outcomes: (1) it identified three particular phenomena which still pose major challenges for REG algorithms aiming to replicate human behaviour, and (2) it provided a platform for the discussion of a number of difficulties that arise for corpus-based evaluation in REG. This resulted in a number of criteria for the design of the two corpora that the work in the remainder of the

thesis is based on.

The three phenomena in the human-produced referring expressions which the tested algorithms were not able to replicate satisfactorily were over-specification, spatial relations, and speaker-specific behaviour. Both GREEDY and the IA were able to generate some of the redundancy that was found in the corpus, but a number of the referring expressions contained more redundant information than either algorithm could produce under any parameter setting, ruling them out as accurate models of human reference behaviour.

Neither GREEDY nor the IA were intended to be able to generate referring expressions that contain relations between entities, but this is exactly what the RA was designed for. Surprisingly, the RA not only failed to generate any of the descriptions contained in the evaluation corpus; the descriptions it did generate seemed more like riddles whose aim was to confuse a listener, rather than helpful attempts at pointing out the target referent. A theoretical appraisal of other approaches designed to handle relations established that none of them would include a relation in the test domain, because a relation was never absolutely necessary to distinguish any of the drawers.

The third area of concern that the experiment highlighted was the observation that people do not all do the same thing in the same situation. In fact, even the same person might describe the same drawer differently under different circumstances. None of the algorithms tested were intended to take such inter- and intra-speaker variation into account, and only very recently have implementations of the IA begun to model speaker-preferences to some degree.

The general issues with corpus-based evaluation that this evaluation experiment uncovered were (1) the tight interdependence between algorithms and the underlying knowledge representation they use, (2) the non-determinism of natural language generation, (3) the question of how to compare algorithms' output to many gold standards, and (4) the domain-specificity of REG algorithms. The discussion of these issues gave rise to the following desiderata for corpus-based evaluation in REG:

1. If corpora are intended to be reused for comparative evaluation of different algorithms, an underlying representation of the domain needs to be provided alongside them.

2. If we want to be confident in any evaluation results based on a REG corpus, this corpus needs to contain as many instances as possible from as many different speakers as possible for each referential scenario. This is true whether an algorithm is evaluated in terms of being able to generate one natural

sounding referring expression, or whether it is tested for its likelihood of being an accurate model of human reference behaviour by checking if it can generate all descriptions in a corpus.

3. If an algorithm's likelihood of being a model of human reference behaviour is evaluated, metrics based on Recall and Precision should be used. In this case, the complete set of descriptions that the algorithm provides for each referential scenario under any parameter setting should be compared to the set of descriptions contained in the corpus for the same referential scenario. If the more application-oriented ability to generate one human-like reference is to be assessed, only one description per scenario should be evaluated. This should be done using a Precision-based metric to test how many of the algorithms' descriptions are contained in the corpus.

4. Algorithms which are tried in one specific domain must not be assumed to be easily adaptable to other domains. Ideally, corpora covering many different types of domains should be made available for testing algorithms' claims of generalisability.

7.1.2 Corpus Collection and Analysis

The outcomes of the evaluation experiment had a direct influence on the corpora that I collected and analysed in Chapter 5. The criteria that I paid specific attention to were (1) that the corpora should make it possible to study human use of spatial relations under conditions where non-relational descriptions could also be used, (2) that each referential scenario (i.e. visual stimulus) should be described by as many participants as possible, and (3) that the corpora would need to be provided with a reusable annotation of the semantic content of the referring expressions contained in them. I assumed that speaker-specific variation and examples of redundant descriptions would occur naturally without being planned for in the corpus design.

The analysis of the two corpora confirmed a number of claims from the literature: First, colour enjoys a position of primacy among the visual properties of at least simple objects. It gets used redundantly much more often than size or location. Second, the use of size, on the other hand, is highly dependent on its usefulness not only in distinguishing the target referent from other objects in general, but from objects sharing the target referent's type in particular. Third, the vertical axis is preferred over the horizontal axes in the use of relations. Objects placed on top of a landmark object are much more likely to be described in terms of the relation to this landmark, than objects placed in front of or next to a landmark.

The outcomes of the corpus analyses in terms of people's use of spatial relations were as follows:

1. People use relations even in situations in which the target referent can be described using only visually inherent attributes such as type, colour and size.

2. There are three types of people: those who never use relations, those who tend to use a relation in every description, and those who vary their behaviour for different stimuli.

3. The use of relations is impacted by the visual salience of the potential landmark object. A landmark object that is visually different from the other objects in the scenes is more likely to be included in a referring expression.

4. The use of relations is impacted by the ease with which the target referent can be described not using a relation. Target referents which can be described with a short non-relational description are less likely to be referred to in terms of their relationship to a landmark than those for which a larger number of visually inherent properties have to be listed.

5. The use of relations is impacted by the discriminatory power of the type of relation that holds between the target and the landmark object. If a second object is present in the scene which stands in the same relation to another object as the target object to the landmark object, the target is less likely to be described using the relations to the landmark than in scenes in which no other object pair is present.

A further important outcome of this chapter was the creation of the two corpora themselves. They are the first sizable collections based on visual domains that contain referring expressions making use of spatial relations between objects. GRE3D7 is by far the largest existing collection of context-free referring expressions to date, and GRE3D3 is comparable in size to the second largest collection of context-free referring expressions for singular targets (the singular portion of the TUNA Corpus).

7.1.3 Corpus-Based Modelling

Having these corpora available made it possible to try a machine learning approach to REG, the subject of Chapter 6. The aims of this chapter were twofold: (1) to characterise people's referring behaviour in terms of scene-based features capturing the target's and the landmark's visual salience, and (2) to attempt to find

similarities between different participant's behaviours despite the influence of the participant-specific preferences revealed by the corpus analysis.

I found that models trained to predict complete referring expressions were only able to accurately characterise the data if the identity of the participants who had produced each instance in the corpus was included as a feature. This confirms the important role that speaker-specific preferences play in the selection of attributes for referring expressions. However, by training trees that predict the inclusion of each individual attribute, I was able to establish that, at this more fine-grained level, more commonality can be found between speakers. Here too, models that had access to speaker identity as a feature were even more successful at characterising the data.

In order to be able to directly compare the behaviour of the individual participants to each other, I trained participant-specific models, one for each attribute. I found that the majority of participants shared each of their attribute-specific models with a large number of other participants. Based on this finding, I advocate the notion of speaker profiles as a more bottom-up approach to attribute selection than the approaches adopted in traditional REG algorithms. Instead of considering attributes as being serially dependent on one another and controlling the outcome mainly by a preference ordering or cost function over the available attributes, a speaker profile consists of a collection of speaker-specific attribute inclusion models which make it possible to consider the attributes independently of each other based on any number of factors that might influence their use. While a speaker's overall behaviour, characterised by the speaker profile, is likely to be highly individualised, the component strategies for each attribute contained in the profile can be shared with other speakers to a much larger degree.

7.2 Future Research Directions

In this thesis, I only considered one-off distinguishing descriptions that have the main aim of identifying the target referent among a set of visual distractors. The assumption was made that the content of the referring expressions at stake here was not influenced by any contextual discourse factors. Simplifying assumptions such as this are necessary in any research endeavour, in order to be able to isolate some causes of an observed phenomenon while excluding others. However, ultimately this approach of studying 'reference in the void', as one might call it, will only be able to explain people's natural referring behaviour up to a certain point. Of course, situations where one person asks another to look at or pass them one object without much previous discourse context, such as the ones simulated in my data

collection experiments, do exist; but even a linguistic context as simple as *Please pick up the...*, the directive that participants in my experiments had to complete, might potentially have an impact on the semantic content of the following referring expression. In an environment with objects of very different kinds, objects that can clearly not be picked up by a human might never be considered as distractors.

Psycholinguistic research using eye-tracking technology has shown that listeners are more likely to consider those objects as distractors for which a proposed action is physically possible (Eberhard et al., 1995; Tanenhaus et al., 1995; Altmann and Kamide, 1999; Chambers et al., 2002). For example, when listening to the sentence *put the ball inside the box*, listeners only looked at container-like objects as soon as they heard the word *inside*, and the stimulus *the boy will eat the cake* made people look at the only edible object in the context, as soon as the word *eat* was uttered. From a listener-oriented perspective, this means that it is not necessary to include attributes in the following reference that distinguish the target referent only from objects that have been ruled out by the linguistic context already. I am not aware of any studies that take a speaker-oriented view on this phenomenon and test whether people take the direct linguistic context into account in the selection of the content for the referring expressions they build. If evidence to this effect can be found, modelling this kind of behaviour would be a natural next step for content selection approaches for REG.

In Chapter 4, I identified the production of over-specified referring expressions as one of the main remaining challenges for REG. Approaches that are based on the Incremental Algorithm are able to generate redundancy to a certain degree, but sometimes people use redundant information in a way that these approaches are not able to mimic. The Graph-based Algorithm, with its fine-grained control parameters, would possibly be able to replicate any referring expression; however, this includes not only any referring expression that a human would also use, but also any other combination of a referent's properties which no human speaker would be likely to use and which would sound bizarre to most human listeners.

At present, we have very little guidance as to how we might be able to separate the good from the bad in this respect. We know that in visual domains people are more likely to include an object's colour redundantly and that the inclusion of an object's size in a description is usually tied more to the usefulness of this attribute in distinguishing it from the distractors around it. The only approaches to REG that I am aware of which have attempted to integrate this information, more or less explicitly, were submissions to the REG evaluation challenges. A number of these approaches, based on the Incremental Algorithm, used preference orders based on the frequency of each attribute occurring in the training data, which implicitly

reflect the tendency to include colour redundantly due to its preference over size as well as other properties in the TUNA Corpus (cf., Hervás and Gervás, 2007; Spanger et al., 2007; Fabbrizio et al., 2008; de Lucena and Paraboni, 2008, 2009). In one of our instantiations of the Graph-based Algorithm, we used a special mechanism to ensure explicitly that colour would be included in every referring expression independently of its discriminatory power (Krahmer et al., 2008; Viethen et al., 2008).

However, the use of redundant properties in relational descriptions has, to my knowledge, never been studied either in psycholinguistics or in computational linguistics. The GRE3D3 and GRE3D7 corpora are well-suited for an investigation into this issue. In both corpora, descriptions of the form ⟨tg_type, relation, lm_type⟩ were always fully distinguishing, but often additional attributes were included for both the target and the landmark object. One open question is, for example, whether it is the relation that has to be considered redundant or whether the additional attributes are the redundant ones. Another interesting question regards specifically the information included about the landmark: are the landmark attributes found in relational referring expressions dependent in some way on the information included about the target referent, or is the landmark distinguished from the context in its own right? An answer to this question might shed light on the cognitive plausibility of algorithms that use a recursive loop for the description of the landmark, and it would help determine whether this recursive loop should be entered with a blank slate, or whether the landmark should already be considered described to some extent via its relation to the target referent.

The referential scenarios for the corpora I presented in this thesis were designed in such a way that relations between entities were never necessary to describe the target referent. I have argued, therefore, that the majority of existing approaches to relational referring expressions would not be able to replicate any of the relational descriptions found in the corpora due to their *a priori* preference for visually inherent attributes over relations. Two potential research directions arise from these circumstances: firstly, an investigation into the possibility of adapting existing relational algorithms to be able to generate relational descriptions even when they are not necessary, followed by an evaluation against the data found in GRE3D3 and GRE3D7; and secondly, an evaluation of existing relational algorithms in a domain in which spatial relations are, in fact, necessary, in order to assess their ability to generate human-like relational descriptions. As we saw for Dale and Haddock's Relational Algorithm in Chapter 4, the in-principle capability of generating relational descriptions in no way guarantees that an algorithm's actual output in a new test domain is satisfactory.

Appendix A

Materials for the GRE3D3 Collection Experiment

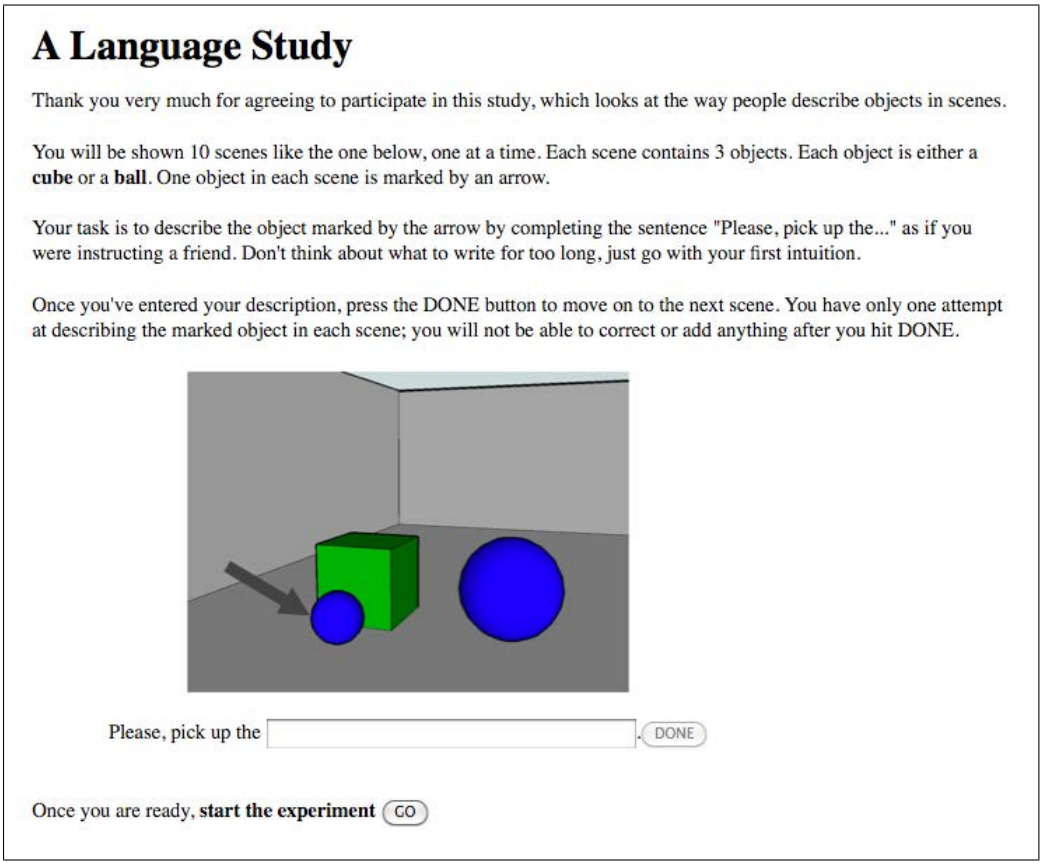

Figure A.1: The instructions for the GRE3D3 collection experiment.

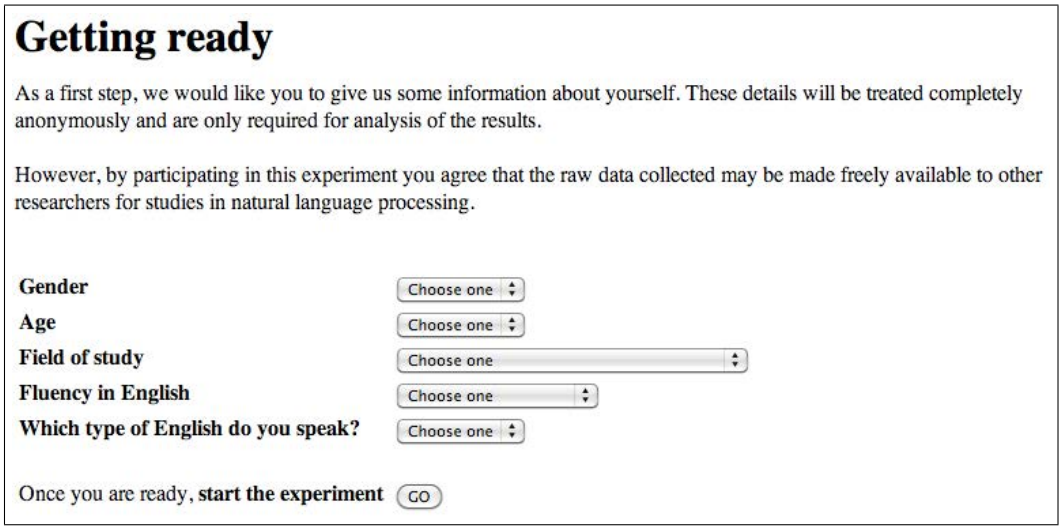

Figure A.2: The form that was used to collect demographic data in the GRE3D3 collection experiment.

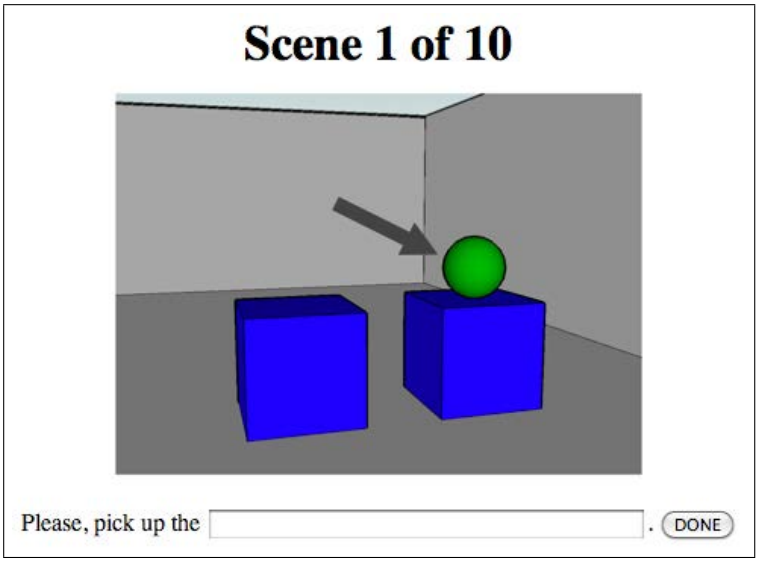

Figure A.3: The screen presenting the first stimulus for the GRE3D3 collection experiment.

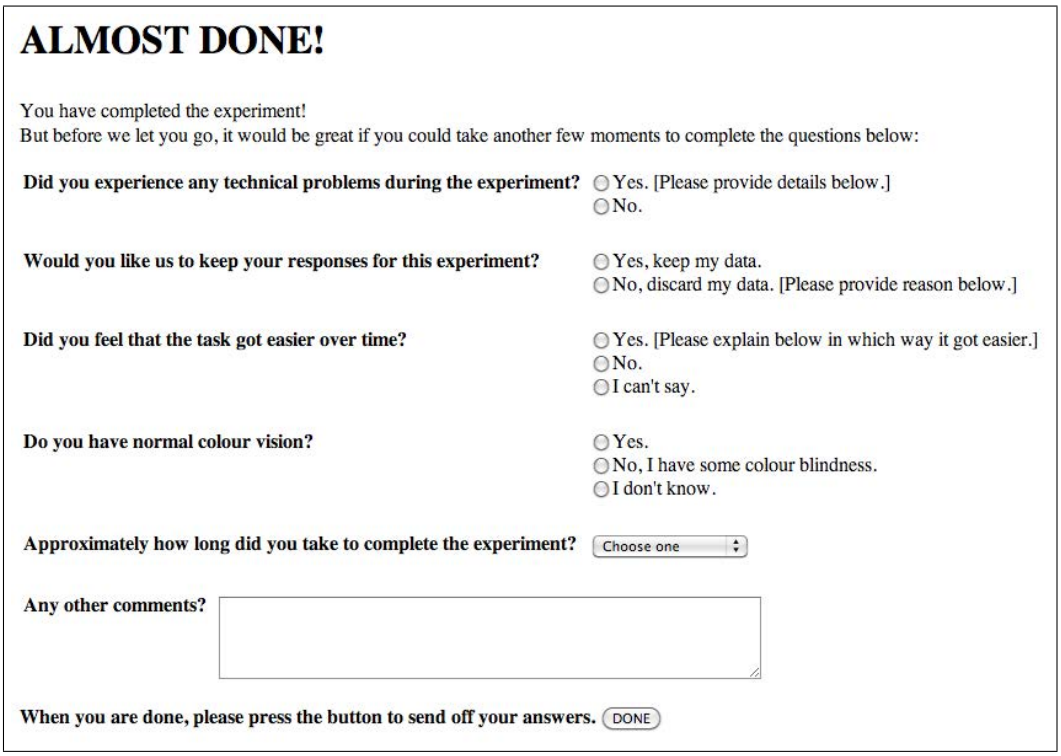

Figure A.4: The exit questionnaire used in the GRE3D3 collection experiment.

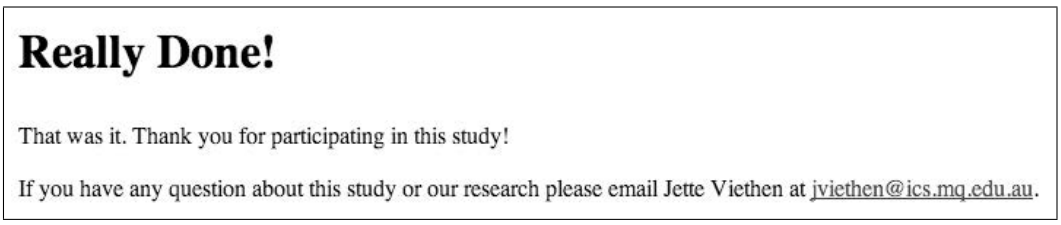

Figure A.5: The final screen of the GRE3D3 collection experiment.

Appendix B

Materials for the GRE3D7 Collection Experiment

B.1 Screenshots of the Experiment

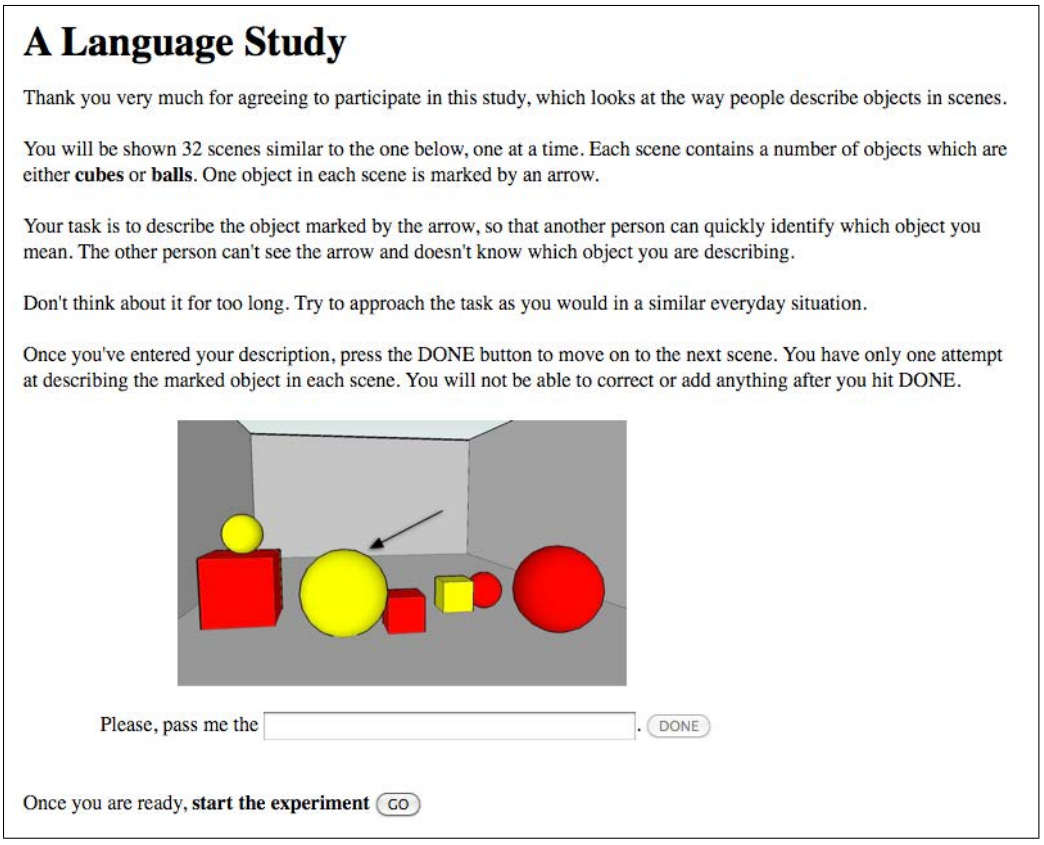

Figure B.1: The instructions for the GRE3D7 collection experiment.

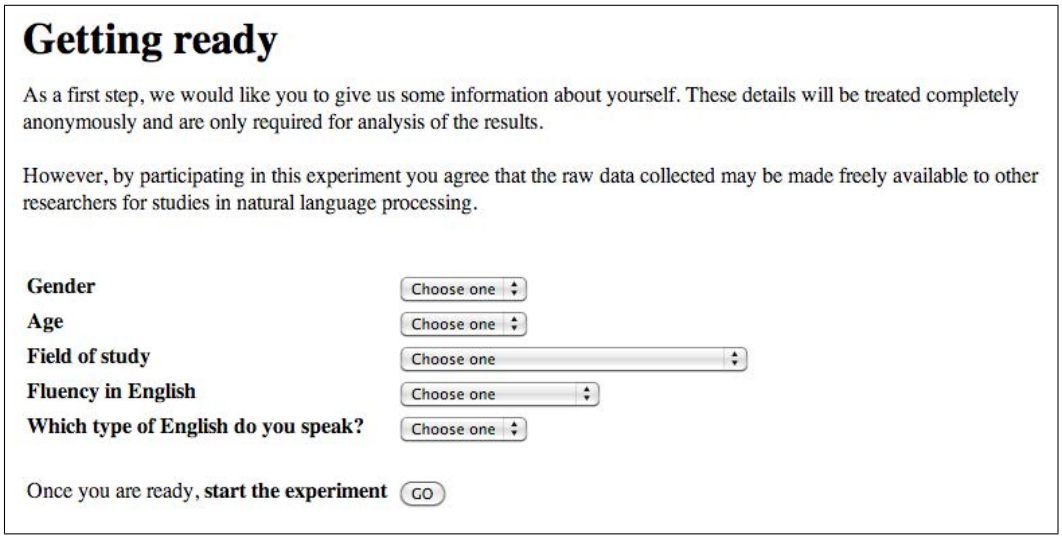

Figure B.2: The form used to collect demographic data in the GRE3D7 collection experiment.

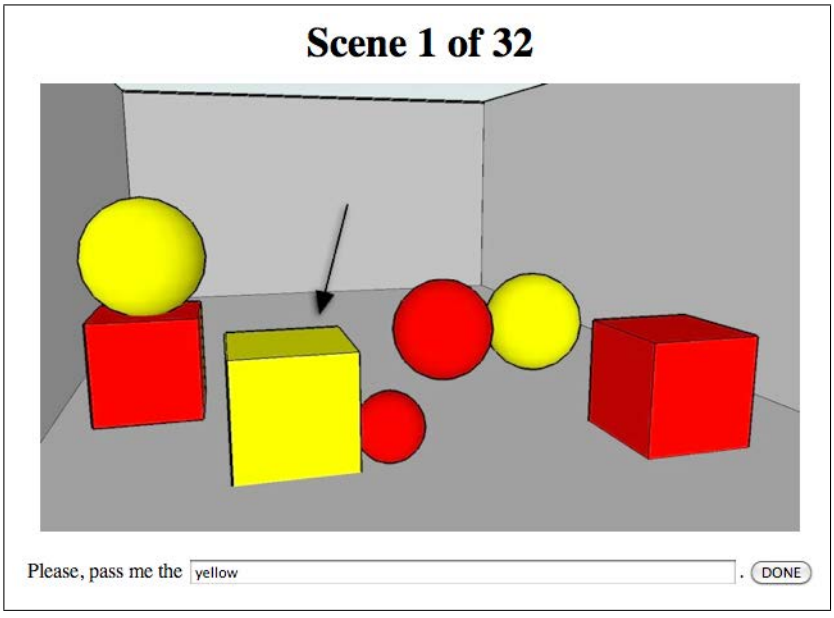

Figure B.3: The screen presenting the first stimulus for the GRE3D7 collection experiment.

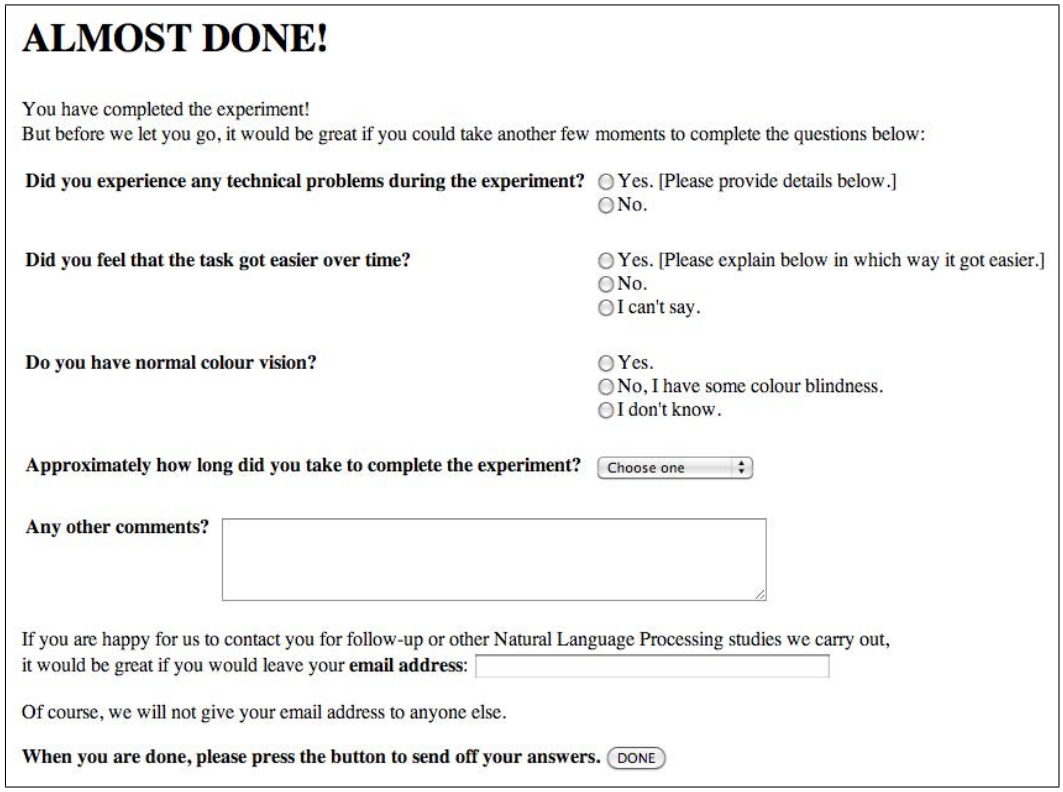

Figure B.4: The exit questionnaire for the GRE3D7 collection experiment.

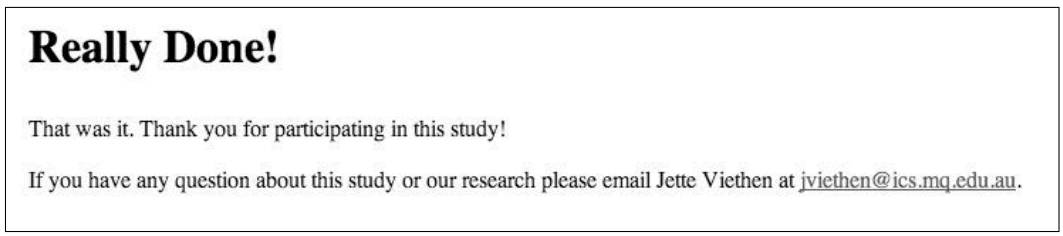

Figure B.5: The final screen of the GRE3D7 collection experiment.

B.2 Filler Scenes

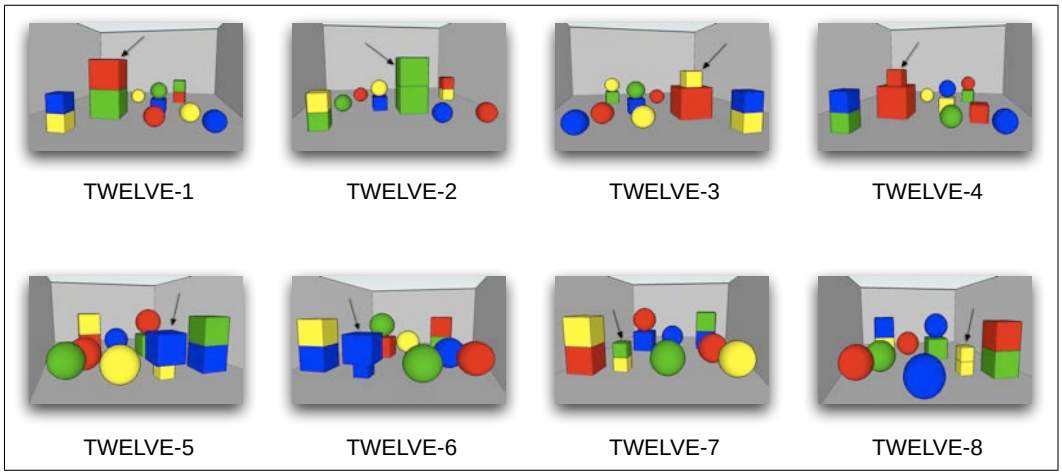

Figure B.6: The eight filler scenes containing twelve objects used in the GRE3D7 collection experiment.

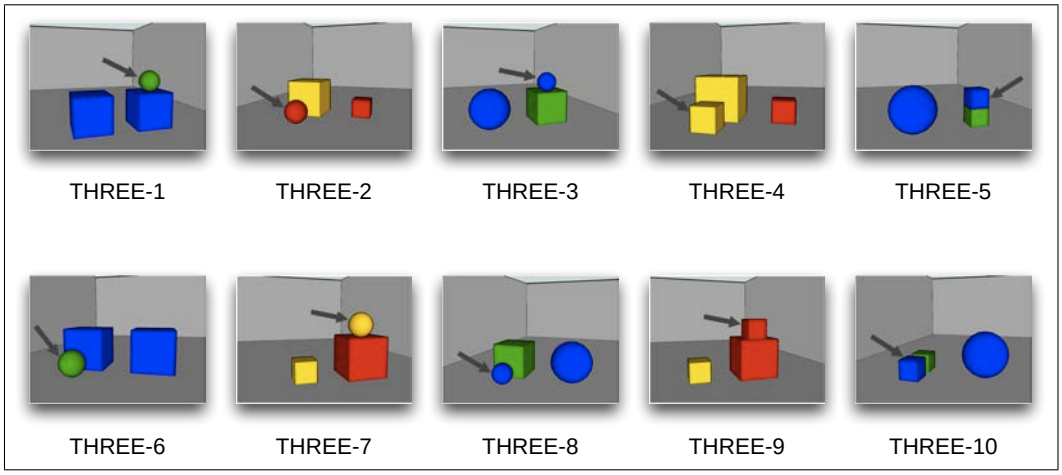

Figure B.7: The 10 filler scenes containing three objects used in the GRE3D7 collection experiment. These are identical with Trial Set 1 used for the collection of the GRE3D3 corpus.

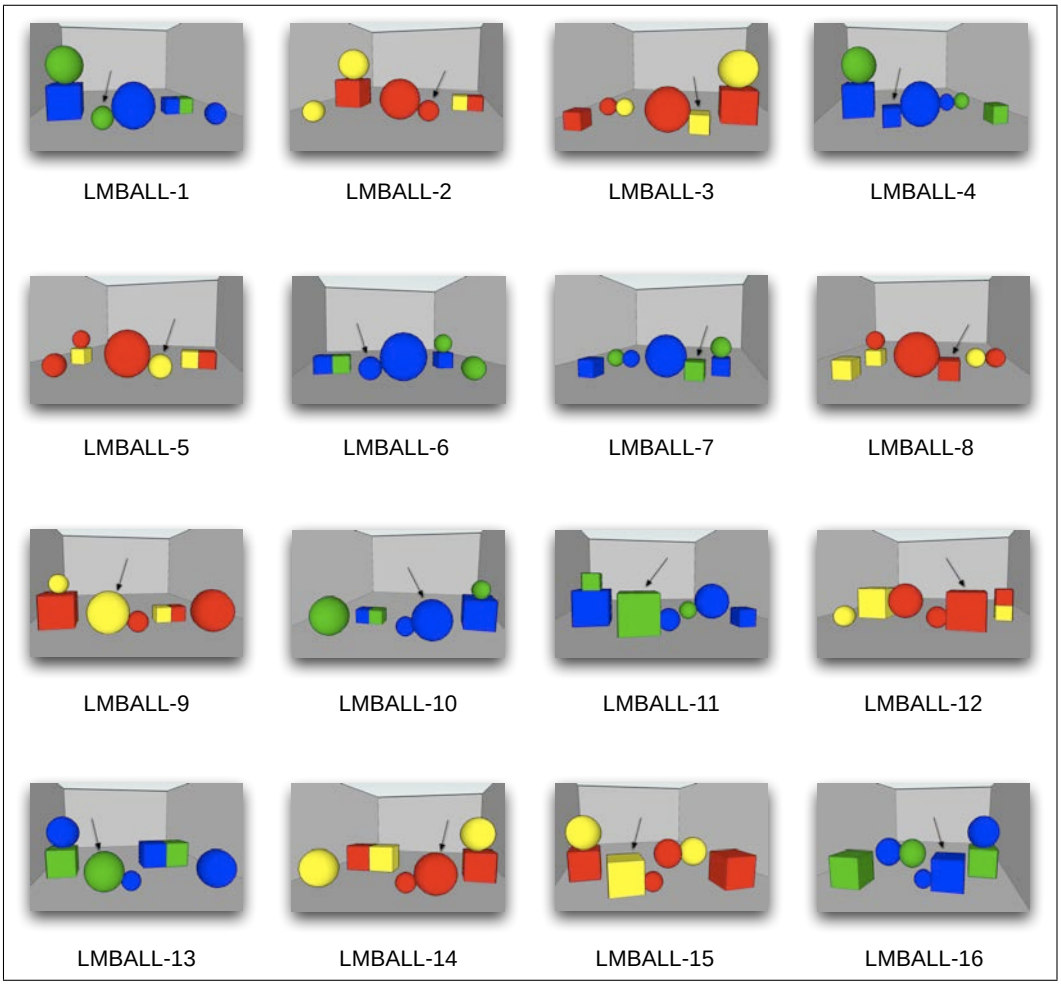

Figure B.8: The 16 filler scenes used in the GRE3D7 collection experiment in which the landmark was a ball instead of a cube

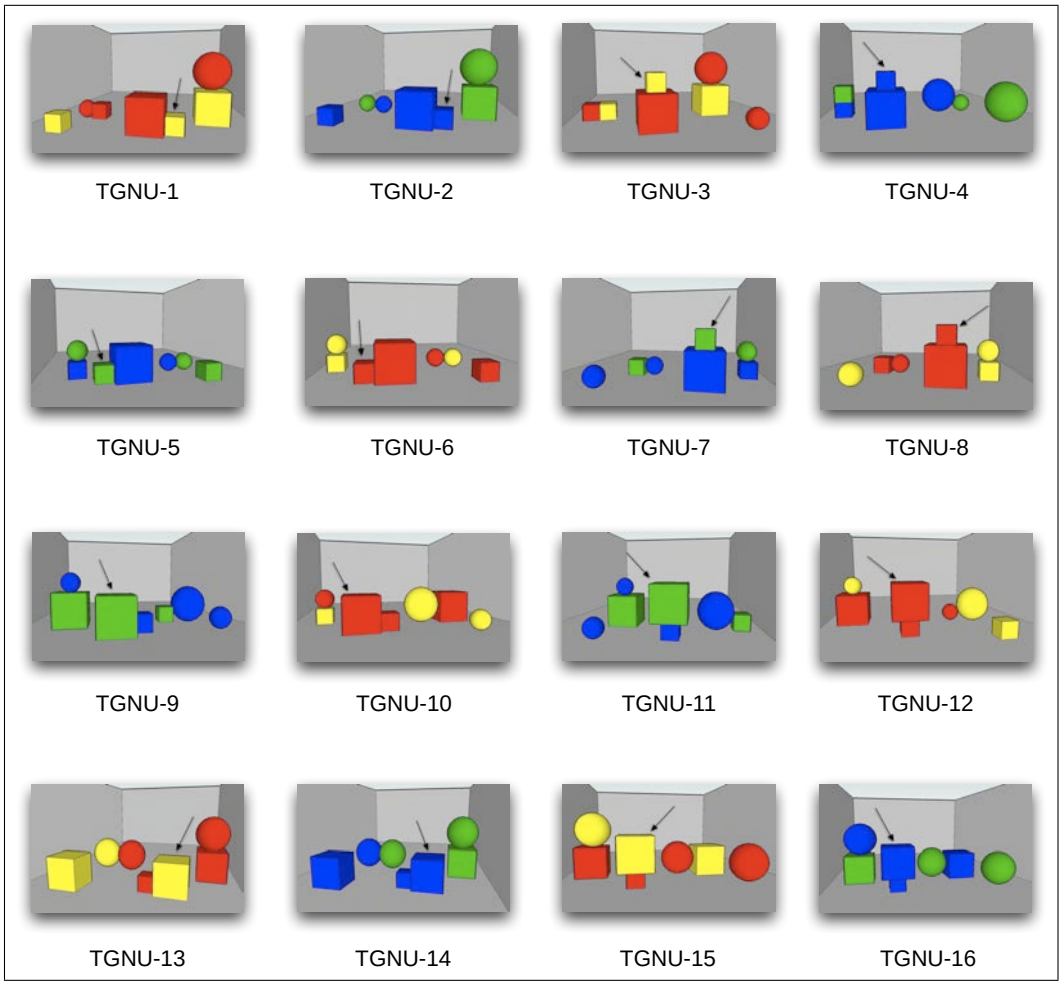

Figure B.9: The 16 filler scenes used in the GRE3D7 collection experiment in which the target is not unique. In these scenes a spatial relation is necessary to fully distinguish the target referent from all other objects.

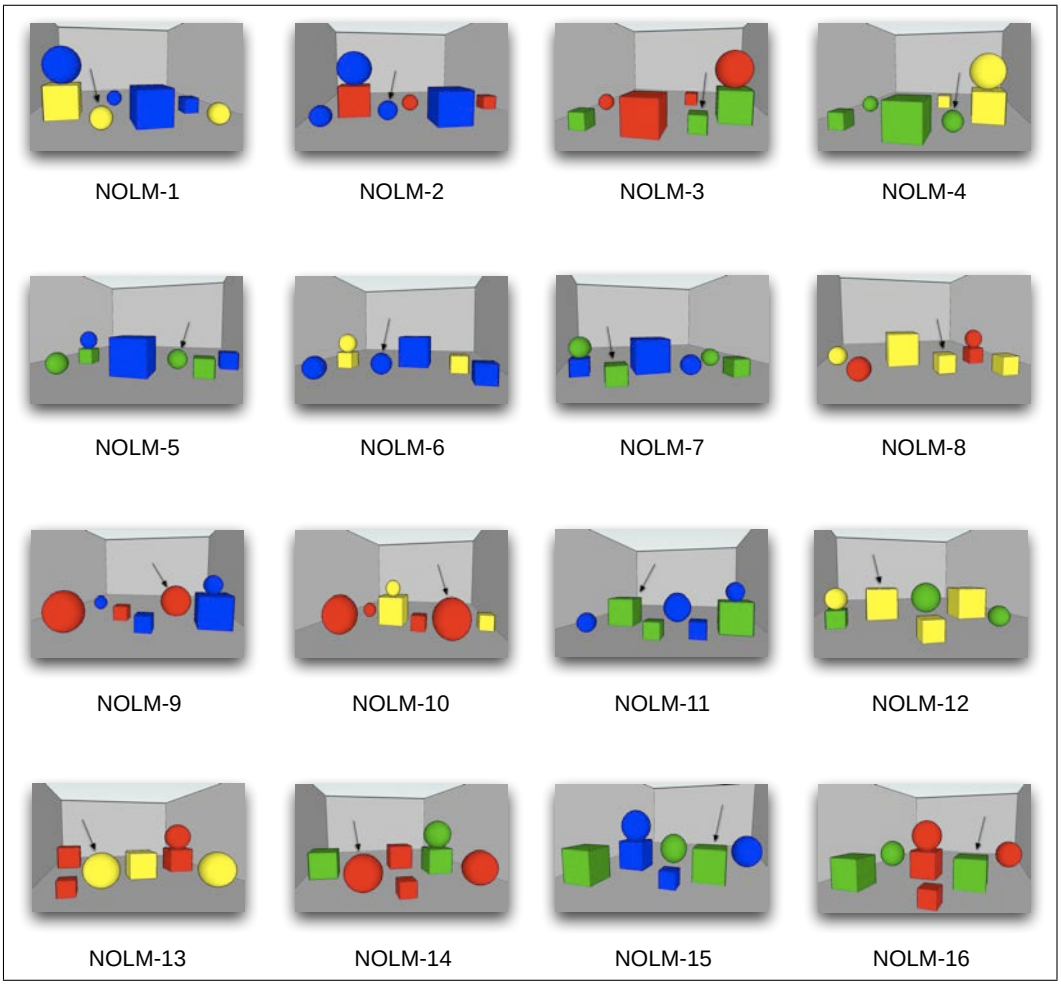

Figure B.10: The 16 filler scenes used in the GRE3D7 collection experiment in which no obvious landmark object exists

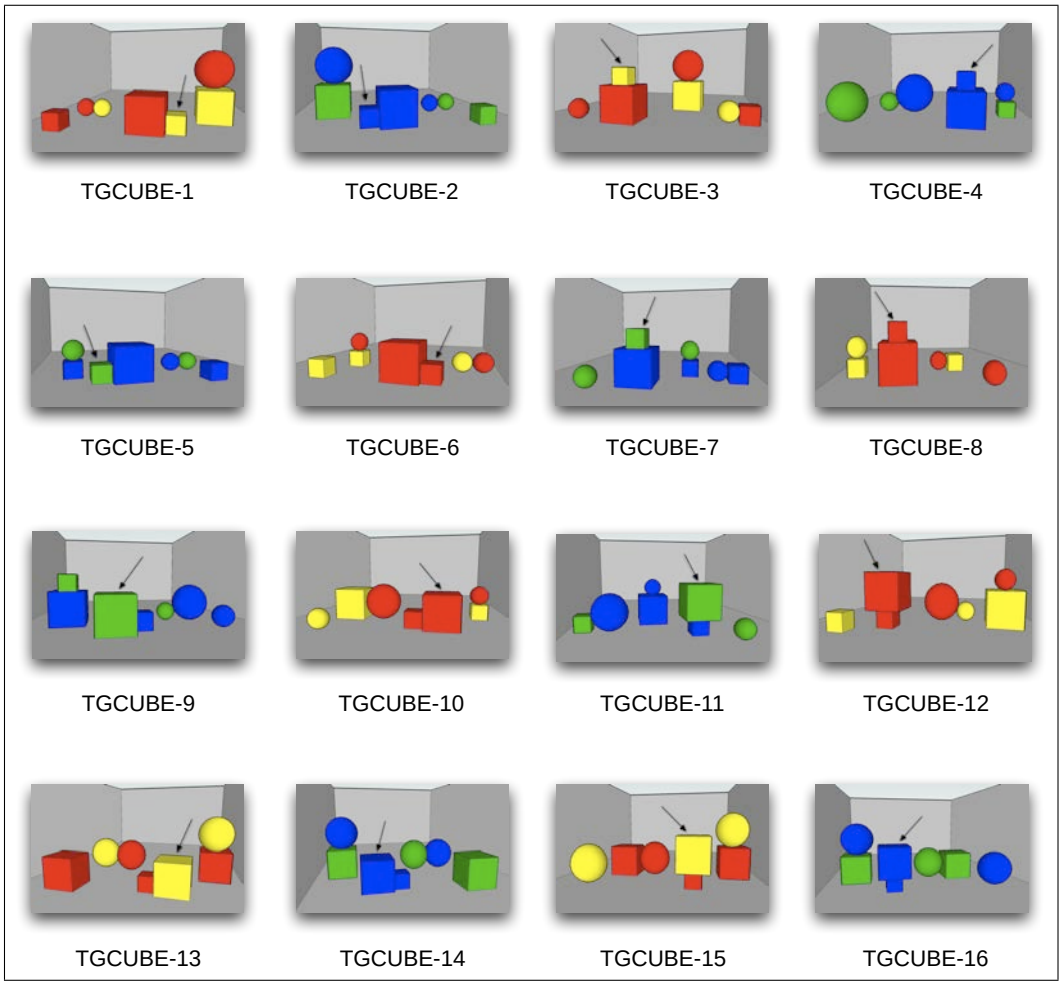

Figure B.11: The 16 filler scenes used in the GRE3D7 collection experiment in which the target is a cube

Appendix C

Tables for Section 6.5.2

	GRE3D3							GRE3D7						
	tg_colour	tg_size	tg_location	relation	lm_colour	lm_size	lm_location	tg_colour	tg_size	tg_location	relation	lm_colour	lm_size	lm_location
Number of speakers sharing each tree	41	40	62	29	35	46	51	273	81	270	215	160	229	280
	15	9	1	15	10	10	9	2	39	2	12	107	35	
	2	4		3	7	2	1	1	26	1	9	5	6	
	2	3		2	2	1	1	1	22	1	6	1	1	
	1	2		2	2	1	1	1	12	1	3	1	1	
	1	2		1	1	1		1	10	1	3	1	1	
	1	1		1	1	1		1	7	1	2	1	1	
		1		1	1	1			6	1	2	1	1	
		1		1	1				6	1	2	1	1	
				1	1				6	1	2	1	1	
				1	1				5		2	1	1	
				1	1				5		...		1	
				1					5		22×1		1	
				1					4					
				1					4					
				1					4					
				1					2					
									2					
									2					
									2					
									...					
									30×1					

Table C.1: Number of speakers sharing each tree. The trees are ordered from most common to least common. Long tails of trees only occurring once are omitted for tg_size and relation in GRE3D7.

error rate	GRE3D3								GRE3D7							
	tg_colour	tg_size	tg_location	relation	lm_colour	lm_size	lm_location	profiles	tg_colour	tg_size	tg_location	relation	lm_colour	lm_size	lm_location	profiles
0	48	43	53	46	56	45	48	21	255	62	238	146	256	238	274	36
1	8	16	9	9	4	17	14	12	18	66	29	68	19	29	4	35
2	2	2		2	2			9	5	74	11	37	4	7	1	48
3	5	2		4	1	1	1	2	1	49	2	18	1	4		38
4			1	2				4		15		6		2	1	46
5								5		12		5				28
6								6	1	2						20
7								1								8
8								1								11
9								1								4
10								1								2
11																2
12																2
total errors	27	26	13	33	11	20	17	147	37	493	57	245	30	63	10	935

Table C.2: Number of participants whose trees had a given error rate.

Appendix D

Publications Related to this Thesis

The following papers and articles have been or will be published about the research reported in this thesis:

Viethen, Jette and Robert Dale (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, 63–70. Sydney, Australia.

Viethen, Jette and Robert Dale (2006). Towards the evaluation of referring expression generation. In *Proceedings of the 4th Australasian Language Technology Workshop*, 115–122. Sydney, Australia.

Viethen, Jette and Robert Dale (2007). Evaluation in natural language generation: Lessons from referring expression generation. *Traitement Automatique des Langues 48(1)*:141–160.

Viethen, Jette and Robert Dale (2007). Capturing acceptable variation in distinguishing descriptions. In *Proceedings of the 11th European Workshop on Natural Language Generation*, 121–122. Schloß Dagstuhl, Germany.

Viethen, Jette and Robert Dale (2008). The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, 59–67. Salt Fork OH, USA.

Viethen, Jette and Robert Dale (2008). Generating referring expressions: What makes a difference? In *Proceedings of the 6th Australasian Language Technology Association Workshop*, 160–168. Hobart, Australia.

Dale, Robert and Jette Viethen (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Natural Language Generation*, 58–65, Athens, Greece.

Viethen, Jette and Robert Dale (2009). Referring expression generation: What can we learn from human data? In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap Between Computational and Empirical Approaches to Reference*. Amsterdam, The Netherlands.

Dale, Robert and Jette Viethen (2010). Attribute-Centric Referring Expression Generation. In Emiel Krahmer and Mariët Theune (Eds.), *Empirical Methods in Natural Language Generation*, no. 5790 in Lecture Notes in Artificial Intelligence, 163–179, Springer, Berlin/Heidelberg, Germany.

Viethen, Jette and Robert Dale (2010). Speaker-Dependent Variation in Content Selection for Referring Expression Generation. In *Proceedings of the 8th Australasian Language Technology Association Workshop*. 81–89, Melbourne, Australia.

Bibliography

Altmann, Gerry T.M. and Yuki Kamide (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73:247–264.

Amoia, Marilisa, Claire Gardent and Stefan Thater (2002). Using set constraints to generate distinguishing descriptions. In *Proceedings of the 7th International Workshop on Natural Language Understanding and Logic Programming*. Copenhagen, Denmark.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson and Regina Weinert (1991). The HCRC Map Task Corpus. *Language and Speech* 34(4):351–366.

Areces, Carlos, Alexander Koller and Kristina Striegnitz (2008). Referring expressions as formulas of description logic. In *Proceedings of the 5th International Natural Language Generation Conference*, 42–49. Salt Fork OH, USA.

Arts, Anja (2004). *Overspecification in Instructive Texts*. Ph.D. thesis, University of Tilburg, The Netherlands.

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

Belz, Anja and Albert Gatt (2007). The Attribute Selection for GRE Challenge: Overview and evaluation results. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 75–83. Copenhagen, Denmark.

Belz, Anja and Albert Gatt (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 197–200. Columbus OH, USA.

Belz, Anja, Eric Kow, Jette Viethen and Albert Gatt (2008). The GREC Challenge: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, 183–191. Salt Fork OH, USA.

Belz, Anja, Eric Kow, Jette Viethen and Albert Gatt (2009). The GREC Main Subject Reference Generation Challenge 2009: Overview and evaluation results.

In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Summarisation (UCNLG+SUM)*, 79–87. Singapore.

Belz, Anja and Sebastian Varges (2007a). Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on Natural Language Generation*, 9–16. Schloß Dagstuhl, Germany.

Belz, Anja and Sebastian Varges (2007b). *The GREC Corpus: Main Subject Reference in Context*. Technical Report NLTG-07-01, University of Brighton, UK.

Beun, Robert-Jan and Anita Cremers (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition* 6(1/2):121–152.

Bohnet, Bernd (2007). IS-FBN, IS-FBS, IS-IAC: The adaptation of two classic algorithms for the generation of referring expressions in order to produce expressions like humans do. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 84–86. Copenhagen, Denmark.

Bohnet, Bernd (2008). The fingerprint of human referring expressions and their surface realization with graph transducers. In *Proceedings of the 5th International Conference on Natural Language Generation*, 207–210. Salt Fork OH, USA.

Bohnet, Bernd (2009). Generation of referring expression with an individual imprint. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 185–186. Athens, Greece.

Bohnet, Bernd and Robert Dale (2005). Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1004–1009. Edinburgh, UK.

Branigan, Holly P., Martin J. Pickering and Alexandra A. Cleland (2000). Syntactic co-ordination in dialogue. *Cognition* 75:B13–25.

Brennan, Susan E. (1996). Lexical entrainment in spontaneous dialog. In *International Symposium on Spoken Dialog*, 41–44. Philadelphia PA, USA.

Brennan, Susan E. and Herbert H. Clark (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:1482–1493.

Brown-Schmidt, Sarah and Michael K. Tanenhaus (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language* 54:592–609.

Brugman, Ivo, Mariët Theune, Emiel Krahmer and Jette Viethen (2009). Realizing the costs: Template-based surface realisation in the GRAPH approach to referring expression generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 183–184. Athens, Greece.

Bryant, David J., Barbara Tversky and Nancy Franklin (1992). Internal and external spatial frameworks representing described scenes. *Journal of Memory and Language* 31:74–98.

Bryant, David J., Barbara Tversky and M. Lanca (2000). Retrieving spatial relations from observation and memory. In Emile van der Zee and Urpo Nikanne (Eds.), *Cognitive interfaces: Constraints on linking cognitive information*, 94–115. Oxford University Press, Oxford, UK.

Byron, Donna, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore and Jon Oberlander (2009). Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation*, 165–173. Athens, Greece.

Caduff, David and Sabine Timpf (2008). On the assessment of landmark salience for human navigation. *Cognitive Processing* 9(4):249–267.

Carletta, Jean C. (1992). *Risk-taking and Recovery in Task-Oriented Dialogue*. Ph.D. thesis, University of Edinburgh, UK.

Carroll, John M. (1980). Naming and describing in social communication. *Language and Speech* 23:309–322.

Chambers, Craig G., Michael K. Tanenhaus, Kathleen M. Eberhard, Hans filip and Greg N. Carlson (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language* 47:30–49.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA, USA.

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

Cohen, William W. (1996). Learning trees and rules with set-valued features. In *Proceedings of the 14th Conference of the American Association for Artificial Intelligence*, 709–716. Portland, Oregon.

Croitoru, Madalina and Kees van Deemter (2007). A conceptual graph approach to the generation of referring expressions. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2456–2461. Hyderabad, India.

Dale, Robert (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 68–75. Vancouver BC, Canada.

Dale, Robert (1990). Generating recipes: An overview of EPICURE. In Robert Dale, Chris Mellish and Michael Zock (Eds.), *Current Research in Natural Language Generation*, 229–255. Academic Press, London, UK.

Dale, Robert (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. Bradford Books, MIT Press, Cambridge MA, USA.

Dale, Robert, Sabine Geldof and Jean-Philip Prost (2005). Using natural language generation in automatic route description. *Journal of Research and Practice in Information Technology* 37(1):89–105.

Dale, Robert and Nicholas Haddock (1991a). Content determination in the generation of referring expressions. *Computational Intelligence* 7(4):252–265.

Dale, Robert and Nicholas Haddock (1991b). Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, 161–166. Berlin, Germany.

Dale, Robert and Ehud Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Dale, Robert and Jette Viethen (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 58–65. Athens, Greece.

Davey, Anthony (1978). *Discourse Production*. Edinburgh University Press, Edinburgh, UK.

van Deemter, Kees (2000). Generating vague descriptions. In *Proceedings of the 1st International Conference on Natural Language Generation*, 179–185. Mitzpe Ramon, Israel.

van Deemter, Kees (2002). Generating referring expressions: Boolean extensions of the Incremental Algorithm. *Computational Lnguistics* 28(1):37–52.

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

van Deemter, Kees and Albert Gatt (2007). Content determination in GRE: Evaluating the evaluator. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 101–103. Copenhagen, Denmark.

van Deemter, Kees and Emiel Krahmer (2007). Graphs and Booleans: On the generation of referring expressions. In Harry C. Bunt and Reinhard Muskens (Eds.), *Computing Meaning*, volume 3, 397–422. Kluwer, Dordrecht, The Netherlands.

van Deemter, Kees, Ielka van der Sluis and Albert Gatt (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation*, 130–132. Sydney, Australia.

Di Eugenio, Barbara, Pamela W. Jordan, R. H. Thomason and Johanna D. Moore (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies* 53(6):1017–1076.

Dice, Lee R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26(3):297–302.

Doddington, George (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145. San Francisco, CA, USA.

Donnellan, Keith S. (1966). Reference and definite descriptions. *Philosophical Review* 75:281–304. [Reprinted in J.F. Rosenberg and C. Travis (Eds.), Readings in the Philosophy of Language, 195–211, Prentice Hall, Englewood Cliffs NJ, USA, 1971].

Eberhard, Kathleen M., M. Spivey-Knowlton, Julie C. Sedivy and Michael K. Tanenhaus (1995). Eye-movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research* 24:409–436.

Edmonds, Philip G. (1993). *A Computational Model of Collaboration on Reference in Direction-giving Dialogues*. Master thesis, Department of Computer Science, University of Toronto, Toronto ON, Canada.

Edmonds, Philip G. (1994). Collaboration on reference to objects that are not mutually known. In *Proceedings of the 15th International Conference on Computational Linguistics*, 1118–1122. Kyoto, Japan.

Eikmeyer, Hans-Jürgen and Elisabeth Ahlsén (1996). The cognitive process of referring to an object: A comparative study of German and Swedish. In *Proceedings of the 16th Scandinavian Conference on Linguistics*. Turku, Finland.

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

Fabbrizio, Giuseppe Di, Amanda Stent and Srinivas Bangalore (2008). Referring expression generation using speaker-based attribute selection and trainable realization (ATTR). In *Proceedings of the 5th International Conference on Natural Language Generation*, 211–214. Salt Fork OH, USA.

Ford, William and David Olson (1975). The elaboration of the noun phrase in children's description of objects. *Journals of Experimental Child Psychology* 19(3):371–382.

Funakoshi, Kotaro, Satoru Watanabe, Naoko Kuriyama and Takenobu Tokunaga (2004). Generating referring expressions using perceptual groups. In *Proceedings*

of the 3rd International Conference on Natural Language Generation, 51–60. Brockenhurst, UK.

Gapp, Klaus-Peter (1995). Angle, distance, shape, and their relationship to projective relations. In *Proceedings of the 17th Annual Meeting of the Cognitive Science Society*, 112–117. Pittsburgh PA, USA.

Gardent, Claire (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 96–103. Philadelphia PA, USA.

Gardent, Claire, Hélène Manuélian, Kristina Striegnitz and Marilisa Amoia (2004). Generating definite descriptions: Non incrementality, inference and data. In Thomas Pechmann and Christopher Habel (Eds.), *Multidisciplinary Approaches to Language Production*, 53–86. Walter de Gruyter, Berlin, Germany.

Gardent, Claire and Kristina Striegnitz (2007). Generating bridging definite descriptions. In Harry C. Bunt and Reinhard Muskens (Eds.), *Computing Meaning*, volume 3, 369–396. Kluwer, Dordrecht, The Netherlands.

Garrod, Simon and Anthony Anderson (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27:181–218.

Gatt, Albert (2006). Structuring knowledge for reference generation: A clustering algorithm. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 321–328. Trento, Italy.

Gatt, Albert (2007). *Generating Coherent Reference to Multiple Entities*. Ph.D. thesis, University of Aberdeen, UK.

Gatt, Albert, Anja Belz and Eric Kow (2008). The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, 198–206. Salt Fork OH, USA.

Gatt, Albert, Anja Belz and Eric Kow (2009a). The TUNA-REG Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 174–182. Athens, Greece.

Gatt, Albert and Kees van Deemter (2005). *Towards a psycholinguistically-motivated algorithm for referring to sets: The role of semantic similarity*. Technical report, TUNA Project, University of Aberdeen, UK.

Gatt, Albert and Kees van Deemter (2006). Conceptual coherence in the generation of referring expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 255–262. Sydney, Australia.

Gatt, Albert, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur and Somayajulu Sripada (2009b). From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Communications* 22(3):153–186.

Gatt, Albert, Ielka van der Sluis and Kees van Deemter (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation*, 49–56. Schloß Dagstuhl, Germany.

Gervás, Pablo, Raquel Hervás and Carlos León (2008). NIL-UCM: Most-frequent-value-first attribute selection and best-scoring-choice realization. In *Proceedings of the 5th International Conference on Natural Language Generation*, 215–218. Salt Fork OH, USA.

Gorniak, Peter and Deb Roy (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21:429–470.

Grice, Herbert Paul (1975). Logic and conversation. In Peter Cole and Jerry L. Morgan (Eds.), *Syntax and Semantics: Speech Acts*, volume 3, 43–58. Academic Press, New York, NY.

Grosz, Barbara J., Aravind K. Joshi and Scott Weinstein (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, 44–49. MIT, Cambridge MA, USA.

Grosz, Barbara J., Aravind K. Joshi and Scott Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Grosz, Barbara J. and Candance L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Guhe, Markus and Ellen Gurman Bard (2007). Adaptation of the use of colour terms in referring expressions. In *Proceedings of Decalog: The 11th Workshop on the Semantics and Pragmatics of Dialogue*, 167–168. University of Trento, Italy.

Guhe, Markus and Ellen Gurman Bard (2008a). Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 2404–2409. Austin TX, USA.

Guhe, Markus and Ellen Gurman Bard (2008b). Adapting the use of attributes to the task environment in joint action: Results and a model. In *Proceedings of Londial The 12th Workshop on the Semantics and Pragmatics of Dialogue*, 91–98. London, UK.

Gupta, Surabhi and Amanda Stent (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, 1–6. Brighton, UK.

Hajičová, Eva (1993). *Issues of Sentence Structure and Discourse Patterns*, volume 2 of *Theoretical and Computational Linguistics*. Charles University, Prague, Czech Republic.

Hardcastle, David and Donia Scott (2008). Can we evaluate the quality of generated text? In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*. Marrakech, Morocco.

Haywood, Sarah, Martin J. Pickering and Holly P. Branigan (2003). Co-operation and co-ordination in the production of noun phrases. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, 533–538. Boston MA, USA.

Haywood, Sarah, Martin J. Pickering and Holly P. Branigan (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science* 16(5):362–366.

Heeman, Peter A. (1991). *A Computational Model of Collaboration on Referring Expressions*. Master thesis, Computer Systems Research Institute, University of Toronto, Toronto ON, Canada. Published as Technical Report CSRI-251.

Hervás, Raquel and Pablo Gervás (2007). NIL: Attribute selection for matching the task corpus using relative attribute groupings obtained from the test data. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 87–89. Copenhagen, Denmark.

Hervás, Raquel and Pablo Gervás (2009). Evolutionary and case-based approaches to REG: NIL-UCM-EvoTAP, NIL-UCM-ValuesCBR and NIL-UCM-EvoCBR. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 187–188. Athens, Greece.

Hopcroft, John (1971). An n log(n) algorithm for minimizing states in a finite automaton. In Z. Kohave (Ed.), *Theory of Machines and Computations*. Academic Press.

Horacek, Helmut (2003). A best-first search algorithm for generating referring expressions. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 103–106. Budapest, Hungary.

Horacek, Helmut (2004). On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation*, 70–79. Brockenhurst, UK.

Horacek, Helmut (2005). Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation*, 58–67. Aberdeen, UK.

Horton, William S. and Boaz Keysar (1996). When do speakers take into account common ground? *Cognition* 59(1):91–117.

Janarthanam, Srinivasan and Oliver Lemon (2009). Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 74–81. Athens, Greece.

Jordan, Pamela W. (1999). An empirical study of the communicative goals impacting nominal expressions. In *Proceedings of the ESSLLI-99 Workshop on the Generation of Nominal Expressions*. University of Utrecht, The Netherlands.

Jordan, Pamela W. (2000a). Can nominal expressions achieve multiple goals?: An empirical study. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 142–149. Hong Kong, China.

Jordan, Pamela W. (2000b). *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, University of Pittsburgh, Pittsburgh PA, USA.

Jordan, Pamela W. and Marilyn Walker (2000). Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 181–190. Hong Kong, China.

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

Kelleher, John (2007). DIT - frequency based incremental attribute selection for GRE. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 90–91. Copenhagen, Denmark.

Kelleher, John and Geert-Jan Kruijff (2005). A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the 10th European Workshop on Natural Language Generation*, 68–74. Aberdeen, UK.

Kelleher, John and Geert-Jan Kruijff (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 1041–1048. Sydney, Australia.

Kelleher, John and Brian Mac Namee (2008). Referring Expression Generation Challenge 2008 DIT system descriptions. In *Proceedings of the 5th International Conference on Natural Language Generation*, 221–224. Salt Fork OH, USA.

Kempen, Gerard and Eduard Hoenkamp (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science* 11:201–258.

King, Josh (2008). OSU-GP: Attribute selection using genetic programming. In *Proceedings of the 5th International Conference on Natural Language Generation*, 225–228. Salt Fork OH, USA.

Koller, Alexander, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore and Jon Oberlander (2010). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference*, 243–250. Dublin, Ireland.

Krahmer, Emiel and Ielka van der Sluis (2003). A new model for generating multimodal referring expressions. In *Proceedings of the 9th European Workshop on Natural Language Generation*, 47–57. Budapest, Hungary.

Krahmer, Emiel and Mariët Theune (1998). Context sensitive generation of descriptions. In *Proceedings of the 5th International Conference on Spoken Language Processing (INTERSPEECH 1998)*, 1151–1154. Sydney, Australia.

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

Krahmer, Emiel, Mariët Theune, Jette Viethen and Iris Hendrickx (2008). GRAPH: The costs of redundancy in referring expressions. In *Proceedings of the 5th International Conference on Natural Language Generation*, 227–229. Salt Fork OH, USA.

Krahmer, Emiel, Sebastiaan van Erk and André Verleg (2003). Graph-based generation of referring expressions. *Computational Lingustics* 29(1):53–72.

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

Landau, Barbara (2003). Axes and direction in spatial language and spatial cognition. In Emilie van der Zee and Jon M. Slack (Eds.), *Representing Direction in Language and Space*, 18–38. Oxford University Press, Oxford, UK.

Levelt, Willem M. J. (1989). *Speaking: From intention to articulation*. MIT Press, Cambridge MA, USA.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10:707–710.

Louwerse, Max M., Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwyneth Lewis, Jie Wu and Megan Zirnstein (2007). Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 1235–1240. Nashville TN, USA.

de Lucena, Diego Jesus and Ivandré Paraboni (2008). USP-EACH: Frequency-based greedy attribute selection for referring expressions generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, 219–222. Salt Fork OH, USA.

de Lucena, Diego Jesus and Ivandré Paraboni (2009). USP-EACH: Improved frequency-based greedy attribute selection. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 189–190. Athens, Greece.

Lyons, John (1977). *Semantics*, volume 2. Cambridge University Press, Cambridge, UK.

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

Metzing, Charles and Susan E. Brennan (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language* 49:201–213.

Miller, George A., Richard Beckwith, Christiane D. Fellbaum, Derek Gross and Katherine Miller (1993). *Five Papers on WordNet*. Technical report, Princeton University, Princeton, N.J.

Nenkova, Ani and Rebecca Passonneau (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL 2004*, 145–152. Boston, Massachusetts, USA.

Olson, David R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review* 77(4):257–273.

Olson, David R. (1972). Language use for communicating, instructing and thinking. In John M. Carroll and Freedle Roy O. (Eds.), *Language comprehension and the acquisition of knowledge*. V. H. Winston & Sons, New York NY, USA.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Technical report, IBM Thomas J. Watson Research Center, Yorktown Heights, NY.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia PA, USA.

Paraboni, Ivandr, Kees van Deemter and Judith Masthoff (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics* 33(2):229–254.

Pashler, Harold (1998). *Attention*. Psychology Press, Hove, UK.

Passonneau, Rebecca (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Passonneau, Rebecca J. (1995). Integrating gricean and attentional constraints. In *Proceedings of the 14th lnternational Joint Conference of Artificial Intelligence*, 1267–1273. Montreal, Canada.

Passonneau, Rebecca J. (1996). Using centering to relax gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech* 39(2–3):229–264.

Pechmann, Thomas (1984). Accentuation and redundancy in children's and adults' referential communication. In Herman Bouma and Don G. Bouwhuis (Eds.), *Attention and Performance X: Control of Language Processes*, 417–431. Lawrence Erlbaum, Hillsdale NJ, USA.

Pechmann, Thomas (1989). Incremental speech production and referential overspecification. *Linguistics* 27(1):89–110.

Pickering, Martin J. and Simon Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2):169–226.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco CA, USA.

Reiter, Ehud (1990a). The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 97–104. Pittsburgh PA, USA.

Reiter, Ehud (1990b). *Generating Appropriate Natural Language Object Descriptions*. Ph.D. thesis, Harvard University, Cambridge MA, USA.

Reiter, Ehud and Robert Dale (1992). A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics*, 232–238. Nantes, France.

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

Reiter, Ehud, Roma Robertson, and Liesl Osman (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* 144:41–58.

Reiter, Ehud and Somayajulu Sripada (2002a). Human variation and lexical choice. *Computational Linguistics* 28(4):545–553.

Reiter, Ehud and Somayajulu Sripada (2002b). Should corpora texts be gold standards for NLG? In *Proceedings of the 2nd International Conference on Natural Language Generation*, 97–104. Harriman NY, USA.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter, Jin Yu and Ian Davy (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence* 67:137–169.

Russell, Stuart J. and Peter Norvig (2003). *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition.

Salton, Gerard and Michael J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York NY, USA.

Schriefers, Herbert J. and Thomas Pechmann (1988). Incremental production of referential noun phrases by human speakers. In Michael Zock and Gérard Sabah (Eds.), *Advances in Natural Language Generation*, volume 1. Pinter Publishers Ltd., London, UK.

Sedivy, Julie C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research* 32(1):3–23.

Siddharthan, Advaith and Ann Copestake (2004). Generating referring expressions in open domains. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 407–414. Barcelona, Spain.

Siddharthan, Advaith and Ann Copestake (2007). Evaluating an open domain GRE algorithm on closed domains. system IDs: CAM-B, CAM-T, CAM-BU and CAM-TU. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 92–94. Copenhagen, Denmark.

van der Sluis, Ielka (2005). *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions.* Ph.D. thesis, Tilburg University, The Netherlands.

van der Sluis, Ielka, Albert Gatt and Kees van Deemter (2006). *Manual for the TUNA Corpus: Referring Expressions in Two Domains.* Technical Report AUCS/TR0705, Computing Department, University of Aberdeen, UK.

van der Sluis, Ielka, Albert Gatt and Kees van Deemter (2007). Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing.* Borovets, Bulgaria.

van der Sluis, Ielka and Emiel Krahmer (2004a). Evaluating multimodal NLG using production experiments. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation.* Lisbon, Portugal.

van der Sluis, Ielka and Emiel Krahmer (2004b). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004)*, 1005–1008. Jeju, Korea.

van der Sluis, Ielka and Emiel Krahmer (2005). Towards the generation of overspecified multimodal referring expressions. In *Proceedings of the Symposium on Dialogue Modelling and Generation at the 15th Annual Meeting of the Society for Text and Discourse.* Amsterdam, The Netherlands.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng (2008). Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.*

Sonnenschein, Susan (1982). The effects of redundant comunications on listeners: When more is less. *Child Development* 56:717–729.

Sonnenschein, Susan (1984). The effects of redundant comunications on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research* 13(2):147–166.

Sonnenschein, Susan (1985). The development of referential communication skills: Some situations in which speakers give redundant messages. *Journal of Psycholinguistic Research* 14(5):489–508.

Sonnenschein, Susan and Grover J. Whitehurst (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research* 11(2):115–126.

Sowa, John F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Cambridge MA, USA.

Spanger, Philipp, Kurosawa Takahiro and Tokunaga Takenobu (2007). TITCH: Attribute selection based on discrimination power and frequency. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 98–100. Copenhagen, Denmark.

Stoia, Laura, Darla Magdalene Shockley, Donna K. Byron and Eric Fosler-Lussier (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation*, 81–88. Sydney, Australia.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard and Julie C. Sedivy (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, New Series* 268(5217):1632–1634.

Tenbrink, Thora (2004). Identifying objects on the basis of spatial contrast: An empirical study. In Christian Freksa, Markus Knauff, Bernd Krieg-Brckner, Bernhard Nebel and Thomas Barkowsky (Eds.), *Spatial cognition IV: Reasoning, action, interaction*, no. 3343 in Lecture Notes in Computer Science, 124–146. Springer, Berlin/Heidelberg, Germany.

Tenbrink, Thora (2005). *Semantics and Application of Spatial Dimensional Terms in English and German*. Technical Report Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition, No. 004-03/2005, Universities of Bremen and Freiburg, Germany.

Theune, Mariët (2000). *From data to speech: language generation in context*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Theune, Mariët, Pascal Thouset, Jette Viethen and Emiel Krahmer (2007). Cost-based attribute selection for GRE (GRAPH-SC/GRAPH-FP). In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 95–97. Copenhagen, Denmark.

Thompson, Henry S., Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands and Cathy Sotillo (1993). The HCRC map task

corpus: natural dialogue for speech recognition. In *Proceedings of the 1993 Workshop on Human Language Technology*, 25–30. Princeton, New Jersey.

Varges, Sebastian (2004). Overgenerating referring expressions involving relations and booleans. In *Proceedings of the 3rd International Conference on Natural Language Generation*, 171–181. Brockenhurst, UK.

Varges, Sebastian (2005). Spatial descriptions as referring expressions in the maptask domain. In *Proceedings of the 10th European Workshop On Natural Language Generation*, 207–210. Aberdeen, UK.

Varges, Sebastian and Kees van Deemter (2005). Generating referring expressions containing quantifiers. In *Proceedings of the 6th International Worskhop on Computational Semantics*. Tilburg, The Netherlands.

Viethen, Jette and Robert Dale (2006a). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, 63–70. Sydney, Australia.

Viethen, Jette and Robert Dale (2006b). Towards the evaluation of referring expression generation. In *Proceedings of the 4th Australasian Language Technology Workshop*, 115–122. Sydney, Australia.

Viethen, Jette and Robert Dale (2007). Evaluation in natural language generation: Lessons from referring expression generation. *Traitement Automatique des Langues* 48(1):141–160.

Viethen, Jette and Robert Dale (2008). The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, 59–67. Salt Fork OH, USA.

Viethen, Jette, Robert Dale, Emiel Krahmer, Mariët Theune and Pascal Touset (2008). Controlling redundancy in referring expressions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

Viethen, Jette, Simon Zwarts, Robert Dale and Markus Guhe (2010). Dialogue reference in a visual domain. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valetta, Malta.

Whitehurst, Grover J. (1976). The development of communication: Changes with age an modeling. *Child Development* 47(2):473–482.

Wilkes-Gibbs, Deanna and Herbert H. Clark (1992). Coordinating beliefs in conversation. *Journal of Memory and Language* 31:183–194.

Witten, Ian H. and Eibe Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco CA, USA.

Yantis, Steven and Howard E. Egeth (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance* 25(3):661–676.

Yu, Jin, Ehud Reiter, Jim Hunter and Somayajulu Sripada (2004). A new architecture for summarising time series data. In *Proceedings of the 3rd International Conference on Natural Language Generation*. Brockenhurst, UK.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort*. Addison–Wesley, Cambridge MA, USA.