COMPUTATIONAL MODELS OF REFERRING

COMPUTATIONAL MODELS OF REFERRING

A Study in Cognitive Science

Kees van Deemter

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by the author using LaTeX.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Deemter, Kees van.
Title: Computational models of referring : a study in cognitive science / Deemter, Kees van.
Description: Cambridge, MA : The MIT Press, [2016] |
Includes bibliographical references and index.
Cartoon illustrations by Regina Fernandes – Illugraphics. All rights reserved.

Identifiers: LCCN 2015038523 | ISBN 9780262034555 (hardcover : alk. paper)
Subjects: LCSH: Reference (Linguistics) | Presupposition (Logic) | Computational linguistics.
Classification: LCC P325.5.R44 D43 2015 | DDC 410.1/835 – dc23
LC record available at http://lccn.loc.gov/2015038523

10 9 8 7 6 5 4 3 2 1

"If you use your finger to point something out to a cat, it will sniff your finger – it won't get the point, so to speak. But human babies point at things before they walk or talk, and apparently with the intention of getting another to focus on the same item (...). This little piece of human behaviour could be seen as the essence and beginning of reference. (...) it lies at the very heart and soul of human language." [Abbott, 2010].

"A simple working system that displays some properties of human memory may suggest other properties that no one ever thought of testing for, may offer novel explanations for known phenomena, and may provide insight into which modifications the next generation of models should include." [Hintzmann, 1990] (writing about models of human memory, and of human cognition more generally)

Contents

Preface

To communicate, speakers and writers need to make it clear what they are talking about. *Reference* anchors their words to people, animals, places, events, and so forth. The act of referring – also known as the production of referring expressions – is thus fundamental to communication. It has been studied so extensively that it might be called the fruit fly of language: just as geneticists have long studied the humble *Drosophila melanogaster* (alias the fruit fly), more than a few cognitive scientists have turned to the seemingly simple phenomenon of reference, hoping that the lessons learned in the study of reference would prove to have wider significance (see section 16.3 for elaboration).

My main aim with this book is to demonstrate that referring is an even more interesting and many-facetted phenomenon than has often been thought, and that computational models of reference offer attractive tools for capturing some of this new-found complexity. To support this claim, the models discussed in this book cover many issues beyond the basic idea of referring to an object, including reference to sets, approximate descriptions, descriptions produced under uncertainty concerning the hearer's knowledge, and descriptions that aim to inform or influence the recipient.

To get the richness of reference across to a broad audience of researchers interested in Cognitive Science is the *primary* aim of this book. I have tried to make each chapter self-contained, presenting algorithms in a uniform way that emphasizes the similarities between them; a glossary of frequently occurring terms and abbreviations is offered at the end of the book. As much as I can, I have written in a manner understandable to a range of cognitive scientists. Part III of the book ("Generating a Wider Class of REs") uses some Formal Logic and set theory, but readers who are less interested in technical details will be able to skip this Part without loosing the thread.

Naturally, the book can be swallowed whole. Other recommended reading strategies include the three in Figure 1: a Psychology and Linguistics path (chapters 1-5, 7, 10.7, and 12-16, skipping most of Part III), a Computer Science path (chapters 1, 4-6, 8, 9, and 12-16, skipping theory and experiments), and a Logic and Philosophy path (chapters 1, 2, 4, 6, 8-11, and 16, skipping psychology, experiments, and Part IV).

My *secondary* aim is to use referring as an example of the computational modelling of a human ability. Computational models of referring belong, first and foremost, to Computational Linguistics, but the study of referring benefits if a range of perspectives is brought to bear, with input from philosophy, experimental psychology, Formal Logic, and Artificial Intelligence. To tell this broader story of reference production as an area of Cognitive Science, and to

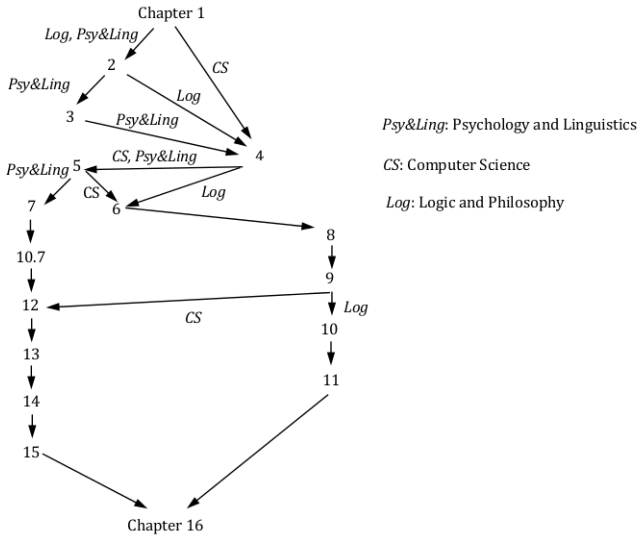


Figure 1

Recommended reading strategies for three types of readers.

discuss the value, and the limitations, of theories and algorithms in this area is the secondary aim of this book.

This monograph focusses on computational models of referring, an area variously known as Generation of Referring Expressions (GRE) or Referring Expression Generation (REG). We are primarily concerned with semantic issues, paying less attention to details of word choice. Moreover, we will focus on aspects of reference that do not rely on what was said before; "one-shot" references take centre stage. The upshot is a book that can be placed within what might be called the Scottish School of reference generation because its focus, though well represented around the globe, coincides with that of a remarkable number of researchers at Scottish universities; the book will focus on research to which my direct colleagues and I have contributed.

Herman Bouma, as a director of Eindhoven's Institute for Perception Research (IPO), nudged me, some 25-30 years ago, towards research that joins up different academic disciplines. For this, and for his wisdom more generally, I dedicate this book to him. At the time, Bouma was a declared advocate of *brevity* – a notion central to the study of referring – and he liked scientific offerings to be as concise as possible. I have tried to take his lessons to heart.

The word "we" can refer to any set of individuals that includes the speaker. This book is full of "we". Sometimes "we" is Albert Gatt, Emiel Krahmer, Roger van Gompel, and I, a gang of four who are endeavouring to create a computational psycholinguistics of reference production. I have learned a great deal from each of the three other gang members, as I have from Ellen Bard, my companion on the REFNET project. Sometimes "we" includes the people in the TUNA project or some of the PhD students I have had the good fortune to supervise. Sometimes "we" includes my colleagues at Aberdeen. I am indebted to all these "we", as colleagues, co-authors, and much more. To shed further light on everyone's role, the introduction to many chapters contains a footnote listing related publications that were jointly authored. Regina Fernandes drew the cartoons for the book: I trust that they will clarify the message, and I hope that readers will enjoy them as much as I do.

This book has benefitted from my teaching at postgraduate courses in Trento (ESSLLI 2002), Tilburg (LOT 2008), Guangzhou (SELC 2010), Harbin (HIT 2010 and 2012), Edinburgh (REFNET 2014), and Aberdeen (NLG Summer School 2015). I am grateful for comments and suggestions from reviewers and the teams at MIT Press and diacriTech. Comments from Robert Dale and Graeme Ritchie have helped me enormously. Other valuable advice was received from Christian Brodbeck, Ronnie Cann, Paul Dekker, Michael Frank, Albert Gatt, Bart Geurts, Roger van Gompel, Matt Green, Frank Guerin, Gerry Hough, Juta Kawalerowicz, Imtiaz Hussain Khan, Alexander Koller, Emiel Krahmer, Wufaldinho Kudde, Roman Kutlák, Vivien Mast, Judith Masthoff, Chris Mellish, Margaret Mitchell, Jeff Z. Pan, Ivandré Paraboni, Paul Piwek, Richard Power, Ehud Reiter, Yuan Ren, Advaith Siddharthan, Melissa Spilioti, and Alice Toniolo. Their help reminds me that "the academic community" can be a community indeed.

Finally, I thank the UK's Engineering and Physical Sciences Research Council, the Cognitive Science Society, the Scottish Informatics and Computer Science Alliance, and the European Science Foundation for supporting the research that underlies this work. The variety of these benefactors speaks to the many aspects of reference.

Kees van Deemter
Aberdeen, January 2016

I FIRST PART: SETTING THE STAGE

1 Aims and Scope of This Book

The Battle of Balaclava is a well-known episode in Britain's history of warfare, culminating in the notorious charge of the light brigade. The charge has been the subject of many paintings and poems.

The battle took place in 1854 during the Crimean War, when Russian armies had captured a large collection of guns from the British troops. The Russians were trying to carry away the guns, which is something the British commander, Lord Raglan, wanted to prevent. From his high vantage point, Raglan was able to oversee the battlefield and decreed, on a sheet of paper carried to the cavalry by a messenger, *"Lord Raglan wishes the cavalry to advance rapidly to the front; follow the enemy and try to prevent the enemy carrying away the guns"*.[1]

The reference "the front" was fatally misunderstood. Lord Raglan intended it as referring to an area known as the Causeway Heights, where Russian troops had gathered and where the guns were being transported (Figure 1.1). The recipient of the message, Lord Lucan, was less well positioned, however, and could see far less of the battlefield. Based on his limited view, the only front that he knew of was an area at the end of a long valley, overlooked from both sides by Russian artillery. Lucan found it difficult to believe that he was asked to cross this valley, because this would expose his men to cannon fire, so he asked the messenger for clarification. The messenger, however, a Captain Nolan, was eager to get on with things and responded irritatedly *"There is your enemy. There are your guns"*, waving vaguely in a direction that was too unclear to be helpful. The rest is history: Lord Lucan followed his commander's order as he understood it: against his better judgment, he led his cavalry through the valley, where Russian cannon were waiting to kill almost the entire brigade (see e.g., [Woodham-Smith, 1954]).

Reference plays a key role in this episode, as when Lord Raglan wrote about "the front". The episode contains many of the issues that will feature in this book. For example, what is being referred to here is not a simple object, but more like a geographical area. Furthermore, the sender and the receiver of the message share much information (where the light brigade is, where the enemy soldiers are), and this allows them to communicate; on the other hand, they have subtly different understandings of some of the facts (e.g., where the guns are, and where there are areas that could be described as a front) and these differences have the potential to compromise communication. Many of

1 The role of communication in this episode was brought to my attention by the legal philosopher Timothy Endicott during the conference *Dealing Reasonably With Blurred Boundaries*, Hannover, Germany, April 2013.

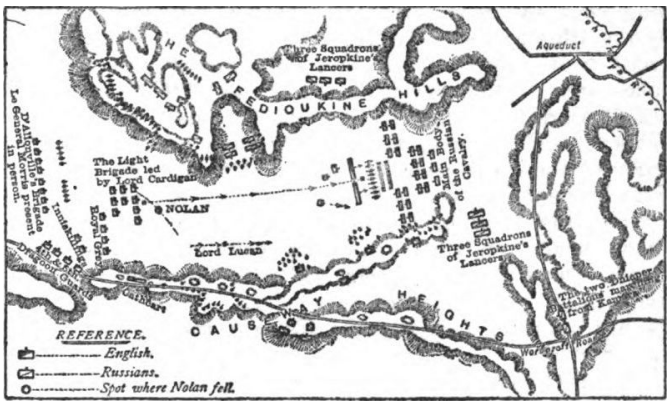


Figure 1.1
The Charge of the Light Brigade at Balaclava, 25 October 1854

the same issues plague everyday conversation. We often give directions, for example, we often don't know (or don't realize) what our hearers know, and we are often misunderstood, albeit usually with less grave consequences.

In the remainder of this chapter, let me explain why reference is important, and what the aims of this book are.

1.1 Aims and Main Thesis

The central thesis of this book is that reference is an even more interesting and many-facetted phenomenon than has often been thought, and that computational models of reference offer attractive tools for capturing this new-found complexity. To support this thesis, the models discussed in this book cover many issues beyond the basic idea of referring to an object, including reference to sets, approximate descriptions, descriptions produced under uncertainty concerning the hearer's knowledge, and descriptions that aim to inform or influence the recipient. Work on these issues, with colleagues in Aberdeen and elsewhere, has been at the centre of my attention for a good ten years, and it is our joint work that will form the core of the book.

My secondary aim is to use referring as an example of the way in which the study of language can be a collaborative affair in which different intellectual approaches come together. I shall show that the study of reference production

benefits if a range of perspectives is brought to bear, with input from philosophy, experimental psychology, Formal Logic, and computing science. To tell this broader story of reference production as an area of Cognitive Science is the secondary aim of this book. Consistent with this aim, I am trying to explain matters in a way that is accessible to researchers across the Cognitive Sciences and Artificial Intelligence.

Reference appears to be a simple idea, yet it is difficult to pin down. John Searle, the philosopher of language, offered a starting point:

> Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular I shall call a referring expression. Referring expressions point to particular things; they answer the questions Who?, What?, Which? [Searle, 1969]

Searle knew that this characterization has some counterintuitive consequences and that it leaves other cases undecided. He therefore opted "to examine those cases which constitute the centre of variation of the concept of referring and then examine the borderline cases in light of their similarities and differences from the paradigms". This is what was done in the survey [Krahmer and Van Deemter, 2012]; it will also be the starting point of the present book, which will end up arguing for a considerable widening of what REG should be.

To see some of the limitations of Searle's definition, consider a quantified noun phrase (NP) such as "No-one". This NP can answer a "Who" question (e.g., Who decided to let the deadline pass?), so, according to Searle's definition, it refers. Yet, it would be difficult to say who the NP refers to. Similar problems affect NPs of the form "only ..", as in "Only one person at a time can pass through this door", which is not about one particular person. Chapter 2 will devote space to a more elaborate – though not entirely conclusive – discussion of these issues.

This book, which covers an area of Natural Language Generation (NLG, e.g., [Reiter and Dale, 2000]), will use the term "referring expression" (RE) loosely, talking sometimes about English NPs and sometimes about the semantic content of such NPs, which linguists might call a Logical Form. Where there is a need to be more explicit, I will use less ambiguous terminology, using terms like "Logical Form". I follow a long tradition of using the word "description" to refer to a wide class of Noun Phrases (NPs), including both definite NPs (which in English tend to use the definite article or a genitive) and indefinite

ones (which tend to use the indefinite article or a bare plural). I follow Robert Dale in using the term *distinguishing description* (and analogously a distinguishing Logical Form) when I want it to be clear that an RE denotes a referent unambiguously [Dale, 1989a]. Referents can be very different kinds of things, which we variously call *objects*, *entities*, or *individuals*.

The book focusses on models of what it is that speakers and writers do when they refer. Because these models take the form of computer programs that generate referring expressions (henceforth, REs), this research area is known as Referring Expressions Generation (REG), or Generation of Referring Expressions (GRE). The present chapter will explain why we study computational models and why we focus on speaking and writing, rather than hearing and reading. But first a few words about reference itself.

Reference is a key component of communication, affecting almost every utterance. For whenever we communicate about specific things, reference anchors our utterances to these things, making it clear that it is them we are talking about. Almost every linguistic subtlety can occur as part of an RE. In fact, practically any sentence of English can be transformed, by means of a simple syntactic operation, into an RE that retains all the complexities of the original sentence. For example, given any sentence S, we can form the complex RE "the idea that S"; presumably if we were able to generate all REs of this form, then we could also generate every sentence S by itself (i.e., without the prefix "the idea that"). If this is true, then being able to generate all REs would mean being able to generate all sentences; apart from issues relating to the supra-sentential structure of text, NLG would be a solved problem.[2] To borrow some technical terms from computational complexity theory, one might say that the (enormous) problem of generating all English sentences *reduces* to the problem of generating all English REs; thus, REG is "NLG complete", because solving REG would mean solving all of NLG. It follows that, in practice, the study of reference production has to focus on a small part of the problem: the problem as a whole is simply too large.

2 Here is another way to derive the same conclusion: any sentence of the form NP VP (noun phrase followed by verb phrase) can be transformed automatically into the RE *The so-and-so who/that VP*. For example, "John walked home because his bicycle broke down" would be transformed into "The person who walked home because his bicycle broke down".

Let's see what this all means in practice. Suppose you were carrying on a conversation in the infirmary of a small zoo, which contains three injured animals, from different countries, having known weights and injuries, as in Table 1.1. We assume that the facts in the table are complete as far as the properties listed are concerned, an assumption known as the Closed World Assumption. For example, lion b has no injuries to his teeth, because this injury is not listed. Which of the listed properties (excluding the identifier, which has been added here merely for convenience) would you employ in an RE? Please try to produce an RE for each of the three animals.

First, consider referent a. Presumably, when you described a, you do not list all its properties. It would suffice to say "The 102kg lion", for example, because the other two animals – the *distractors*, as we shall say – have different body weights. On the other hand, the difference with the weight of b is so small as to be practically negligible. Maybe you preferred to say "The Kenyan lion", unless you felt that the medical nature of your visit makes it more relevant to say "The lion whose teeth are injured". This, however, gives the mistaken impression that the animal has no other injuries, so maybe it would be better to say "The lion that has injured paws and teeth", or "The doubly injured lion". None of all these REs is minimal, in the sense of offering no information that is surplus to the requirement of identifying the referent uniquely: they contain the property "lion", which is unnecessary given that the other properties suffice to rule out both distractors. Note, finally, that even an incorrect description can sometimes work well: if you said "the Tanzanian lion", getting the animal's origin slightly wrong, your audience would probably understand you correctly, though the seed of future misunderstandings would have been sown.

Next, let's turn to b and c. As for b, you could focus on the country of origin, saying "The Chinese lion". As for c, you can simply say "the tiger", though this time the animal's weight stands out enough to make it worth highlighting, as in "The tiger that weighs over 300kg", or even "The huge tiger", suppressing detail. Alternatively, you might prefer to mention what's most relevant in the context of the infirmary, saying "The tiger with the injured back".

Your referential options become even more varied if your task is to refer to a and c together: you could mention them individually. If you wanted to be brief, you could say "The Kenyan and the tiger" (or "The Kenyan animal and the tiger", because the noun makes the reference more felicitous), though the asymmetry between the two halves of this RE might lead you to opt for the

IDENTIFIER	SPECIES	ORIGIN	WEIGHT	INJURIES
a	*lion*	*Kenya*	$102kg$	*paws, teeth*
b	*lion*	*China*	$100kg$	*paws*
c	*tiger*	*China*	$310kg$	*back*

Table 1.1
Information shared by speaker and hearer

lengthier "The Kenyan lion and the Chinese tiger". If animals from other countries were added, you could find yourself aggregating species and geographical regions, as in "the Asian mammals".

Each of these expressions succeeds in singling out the target referent, by saying something that's true of the referent but false of the other animals. Yet the choice between these expressions matters. Some are more fluent, or clearer, than others. Moreover, each RE highlights a subtly different aspect of the animals in question. Understanding all these differences is the ultimate aim of researchers who construct computational models of referring.

1.2 Reference in Practical Applications of Computing

Reference affects not only communication between people, but some of the oldest applications of computing as well. It will be instructive to take a quick look at a few of them, because they form the context in which many REG algorithms were (and still are) designed. For although these algorithms can be seen as computational models of a human ability (see the Epilogue of this book), they are also practical tools.

In database management, reference lies at the heart of *entity resolution* [Newcombe et al., 1959], [Elmagarmid et al., 2007]. One version of this problem is to decide, for two items, in two different databases, whether they represent the same real-world entity. In [Croitoru et al., 2011], a librarian wants to enter the authorship of a book into a database. If another book in the library was written by the same person, then the new book should be entered into the database in such a way that both items are shown to have the same author. But two different authors may have the same name. To decide whether two books have the same author can be difficult. For example, if someone has written a book on mathematics in 1990, is she likely to have written a book on biology in 2015 (Figure 1.2)?

Figure 1.2
Entity resolution: Is this the same Jane Smith or a different one?

Another source of reference problems is *Information Extraction* (IE), where structured information is gained from a text, by means of a limited kind of text interpretation. Suppose the input to IE is a text about corporate mergers:

> "Bridgestone Sports Co. said Friday that it has set up *a joint venture* in Taiwan (...) *The joint venture*, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990" [Grishman and Sundheim, 1995], see also [Jurafsky and Martin, 2009].

A typical aim is to fill the slots in a template, for example:

> Initiating company: *Bridgestone Sports Co.*
> Joint venture: *Bridgestone Sports Taiwan Co.*
> Starting date: *January 1990*

This involves a number of tasks that have reference at their heart: *Named Entity Recognition* means figuring out the intended referent of names like "Bridgestone Sports Taiwan Co." This is challenging because companies may

be known under different names, and similar names may denote different entities (Bridgestone Sports Co. is not Bridgestone Sports Taiwan Co.). Another task is *coreference* resolution, that is, determining which parts of a text are about the same things. This can mean, for example, finding out that the joint venture introduced in the opening sentence is identical to the one that started up in January 1990. IE programs do this by working out that the two italicized NPs ("a joint venture" and "the joint venture") refer to the same entity.

Reference is equally crucial in *Natural Language Generation* (NLG). An example is the new "Robot Journalism", in which newspaper articles are constructed wholly or partly by computer programs: such articles routinely refer to people, companies, and so on. Another example are intelligent interfaces, such as SIRI, Apple's personal assistant for mobile phones. SIRI conducts multi-turn conversations, adjusting references to information assembled on the fly. If the user says "Meet with Jamie for coffee at 2", the system may respond "OK, I scheduled your meeting with Jamie Chen at 2PM today", declaring its interpretation of references to a person and a time. SIRI works with panache yet, at the time of writing, struggles to interpret the REs produced by users.

Data Anonymization [Raghunathan, 2013], finally, is the opposite of REG: whereas the former tries to make it easy for a recipient to identify something or someone, the latter tries to make this hard. For example, a medical database may contain information about patients suffering from certain conditions, the operations that they underwent, and so on. Often it is a legal requirement that the database does not give away the identity of any patient. To meet this requirement, it does not suffice for the patients' names to be omitted: a combination of their date of birth and current address, for example, would give away their identity just as effectively. To figure out what information the database should be permitted to contain is a problem that echoes REG in many ways.

1.3 Computational Models of Reference Production

This book will focus less on applications and more on the challenge of understanding how reference works, using computational models as our tool. Its subject may be seen as an area of *theoretical* NLG, as opposed to the type of NLG that sees the construction of practical systems as its be-all and end-all.

Psycholinguists study reference using an approach based on experiments with human participants (see chapter 3). Theoretical linguists and philosophers of language analyse how definite descriptions contribute to the meaning of the

sentences in which they occur (chapter 2). The present book aims to demonstrate that an *algorithmic* perspective has much to add to, as well as learn from, these accounts because the construction of algorithms forces one to be explicit where others can wave their hands. This point is worth explaining in more detail, because its relevance extends far beyond the study of reference.

A growing number of researchers are becoming aware that algorithms can bolster theory [Poesio et al., 2004] [van Deemter et al., 2012c]. For example, in a theory article, one might read that a speaker should identify the intended referent using "one or more of its contextually salient properties", where it is left unspecified what the context is, how the salience of a property (section 4.9) should be measured, and how the properties in question are selected. Although it is legitimate to leave such matters unspecified, and letting experiments concentrate on clear cases, other issues come to the fore when we are forced to be more specific. Algorithms can help us make our ideas precise.

Reference is associated with two different processes. The first is known by psycholinguists as *production* of language, whereas computational linguists usually speak of *generation*. The second process concerns the role of reference in hearing and reading; this is known as *comprehension*. Both processes have been modelled algorithmically, with computer programs taking the place of a person. Generation algorithms take the domain, and an intended referent in it, as their input and deliver an RE as their output. Interpretation algorithms take the domain and an RE, and deliver a referent as their output.

Until recently, language production and generation were studied far less often than interpretation, but this is changing, based on a growing body of accepted methods [Levelt, 1989], [Reiter and Dale, 2000], [Bateman and Zock, 2002], [Krahmer and Theune, 2010]. The computational generation of REs has been studied especially in recent years (see [Krahmer and Van Deemter, 2012] for a survey). Generation focusses our attention on the issue of *expressive choice*: for example, there are typically many ways in which a given referent can be identified in principle, yet some of these come much more naturally to speakers and some are more useful to hearers. We shall see that expressive choice is a rich and rewarding research topic; in fact, there are many cases (studied particularly in Parts III and IV of this book) in which it seems a miracle how any speaker is able to choose a halfway decent RE from among the myriad of awful ones. As we shall also see, algorithms in this area are increasingly regarded as interesting – though imperfect – models of human reference production.

Stages	Psycholinguistics	Computational Models	Task
Stage 1	Conceptualization	Content Determination	What to say
Stage 2	Formulation	Linguistic Realization	How to say it
Stage 3	Articulation	Speech Synthesis	Saying it

Table 1.2

Three stages often distinguished in human speech production, with their approximate computational counterpart. (For the process known as self-monitoring, see text.)

Perhaps the deepest theoretical question on which the computational study of reference production touches is the question of rationality in language use. Some models of reference production have been motivated by an appeal to rationality – in the form of the Maxims, for example (chapter 4), or in terms of making the referent easy to find by an idealized hearer – whereas other models are best seen as heuristics that work well in some cases but that can also misfire (cf., [Kahneman, 2012]). Once again, the Battle of Balaclava is a reminder of the irrationality of human behaviour: Lord Raglan, the author of the disastrous message, should have realized that the recipient could not know where the guns were; Captain Nolan, when asked for clarification, should have realized that a terrible error was about to be committed. And yet they did not.

1.4 Determining the Information Content of an RE

Psychologists believe that human speech production involves three consecutive stages: before speaking, a speaker first has to decide what to say, then how to say it. Details of what each stage comprizes can differ, and so can the terminology that is used [Garrett, 1984], [Levelt, 1989], [Dell et al., 1997], [Levelt et al., 1999], [Vigliocco and Hartsuiker, 2002]. The decision what to say, however, is frequently referred to as *conceptualization*; the decision how to say it is sometimes known as *formulation* (this includes lexical access and planning of the surface structure of the utterance); the process that produces actual speech is called *formulation* (Table 1.2).

Computational models tend to mirror these three stages: models of Stage 1 are known as *Content Determination*; models of Stage 2 as *Linguistic Realization*; and models of Stage 3 as *Speech Synthesis*. Similar to psycholinguistic models, the output of Content Determination is usually fed into Linguistic Realization (although alternative architectures exist, which interleave the two

components, see section 6.6), whose output is fed into Speech Synthesis. When components are daisy-chained in this way, we obtain what is sometimes known as the standard NLG pipeline [Reiter and Dale, 2000]. Elaborations and variants of both pipelines have been proposed, but these three stages can usually be discerned. In psycholinguistics, a fourth component is often added, namely, *self-monitoring*, in which a speaker examines the output of one of the three stages; subsequently, the output may be modified ([Levelt, 1989], chapter 12). Interesting though it is, computational models have so far had little to say about this fourth stage, and we shall only rarely refer to it (see sections 8.5 and 12).

This book will focus on computational models of Conceptualization. Specifically, because we are interested in reference, the book will focus on the conceptualization of *referring expressions*, an area known as "Content Determination for REG", where the last two words are omitted when they can be inferred from the context. In practice, this means that our discussion will focus on the semantic core of reference production, often disregarding details of syntactic structure and word choice; this explains the dearth of syntax trees in the book.

In the simplest cases, the output of Content Determination can be a set of properties, all of which hold true of the referent. For example, Content Determination can produce a Logical Form such as $\{car, blue\}$, which will be understood as the logical conjunction of the two properties. In Part III of the book, we shall be needing increasingly complex Logical Forms, in order to match the content of more elaborate REs. At that stage, the convenient shorthand of sets of properties will no longer serve us, so we shall be using more elaborate notations (e.g., from Description Logic) to express negation, relations, quantifiers, and so on. At that point, the term "Logical" Form will no longer feel like a misnomer.

I am focussing on Content Determination because once a set of properties has been selected, expression of these properties by means of English words is a task that is not specific to REG: a very similar Linguistic Realization task is relevant to the generation of all Noun Phrases (NPs), and Verb Phrases too. Linguistic Realization requires that suitable words are found to express each concept, and to arrange these in a linguistically appropriate order – it is usually much better to say "large blue car" than "blue large car" – and to decide which concepts are realized as pre-modifiers, and which as post-modifiers, for example; realization solutions that work well for REs include [Malouf, 2000] and [Mitchell et al., 2011a]. I will only discuss Linguistic Realization of REs where a bad realization decision threatens to create referential ambiguity – the cardinal sin of reference production (section 8.8).

A considerable amount of research is invested in finding out how REs are realized in speech, as a function of the context in which they occur [Bard and Aylett, 2004], and in combination with gestures [de Ruiter, 2000]. A considerable proportion of this work has reference as its focus, and much of it is linked with speech science and embodied conversational agents (e.g., [Cassell et al., 2000]). These issues will not be covered here; doing justice to them would deserve a separate book.

Work on reference production is often conducted under the pretence that findings in this area must always be language independent. This is not a thoroughly tested assumption – to put it mildly – but it is more plausible if one focusses on Content Determination (i.e., on concepts rather than words).[3] Future work will almost certainly lead to more nuanced insights but, for better or worse, we will be focussing primarily on English REs.

In view of my decision to focus on Content Determination, it will be convenient to commit a slight abuse by using the term "referring expression" (and the abbreviation RE) ambiguously: on some occasions it will denote an actual linguistic expression (usually an NP), but on others it will denote a Logical Form that can be expressed by different English NPs.

1.5 Focus on Speakers or Hearers?

The focus on this book will be on computational models of a human ability. In the Epilogue (section 16.1), we shall ask what we can learn from looking at this enterprize through the prism of Computational Cognitive Modelling. A few things are worth saying in advance.

Not long ago, researchers in Natural Language Generation were content to produce "felicitous" output without saying clearly what this means. Today, we tend to be clearer about our aims, and consequently about the manner in which a model should be tested. As it happens, our aims can differ: some studies ask what RE a speaker would be most likely to say, whereas others ask what REs are most effective in terms of their utility for hearers or readers. In both cases the outcome is a model. Where the work is motivated by practical applications

3 Occasionally, languages are compared. An example is a set of studies reported in [van Gompel et al., 2012], where our interest in the TYPE attribute led us to compare REs produced in English and Dutch (where we had reason to expect a smaller preference for types); as it happened, any differences turned out to be subtle. See also [Koolen et al., 2009]; and [Khan, 2015] for reference in Arabic.

(e.g., section 1.2), the latter types of models (which focus on utility for hearers) are more relevant than the former (which aim to mimic speakers). Where the work is motivated by theoretical considerations, the emphasis can be on either type of model, though the former type is studied more often than the latter.

The book will try to do justice to both types of research. Models that mimic speakers will take centre stage in the first two Parts, because the bulk of the work on simple REs (what I call the classic REG task) has predominantly taken this perspective. Parts III and IV will switch back and forth between the two perspectives. Occasionally the two perspectives will merge, because in trying to find an effective way to speak, it can be useful to seek inspiration from human language production. Conversely, a model of what is optimal for hearers can shed light on speakers' failure to communicate optimally.

1.6 Referring in One Shot

Linguistic context can affect reference, for instance, by facilitating the use of personal pronouns and demonstrative NPs. A vast amount of work in linguistics ([Ariel, 1988], [Gundel et al., 1993], [Kamp and Reyle, 1993]), psycholinguistics [Arnold, 2008], and computational linguistics (e.g., [Mitkov, 2002], [Stoia et al., 2006], [Siddharthan et al., 2011]) is devoted to these phenomena. To do justice to them would have distracted attention away from what I take to be the core of the phenomenon of reference. There are systematic reasons also for leaving them aside: the modelling of contextual phenomena requires mechanisms that tend to be strongly language dependent, for example, because different languages have different sets of pronouns (e.g., with or without gender; with or without null pronouns).[4] In fact, these mechanisms are so bound up with specific words ("he", "it", "herself", "those", etc.) that our focus on Content Determination (section 1.4) would have been very difficult to uphold.

In recent years, increasing attention is being devoted to the fact that speakers and hearers can actively collaborate to refer (see section 3.6): the speaker may start out with a description that is not quite clear enough, but further interactions with the hearer may clinch the deal. Psycholinguists have rightly asked attention for these phenomena (e.g., [Clark and Wilkes-Gibbs, 1986a], [Di Eugenio et al., 2000]).

4 See also [Piwek, 2008] on the differences between demonstratives in English and Dutch.

Clark and Wilkes-Gibbs explained the matter by defining a class of REs that meets the following conditions:

> The RE is expressed with a proper name, a definite description, or a pronoun;
>
> The speaker uses the RE intending the addressee to be able to identify the referent uniquely against their common ground;
>
> The speaker satisfies her intention by the issuing of that NP; and finally,
>
> The course of the process is controlled by the speaker alone.

The mistaken idea that all REs function in this way was called "the literary model of definite reference". The role of collaboration is now so well appreciated that a new orthodoxy appears to be taking hold; at its boldest, it asserts that conversation is the only legitimate site of language use, and that collaboration is what reference is all about. There is no doubt that reference can involve collaboration and negotiation, just like a sales transaction can involve bargaining. Yet there are many situations in which collaboration is not necessary or not possible (just as we can buy something without bargaining). Examples occur when we write an email or a journal article. Lord Raglan's reference to "the front", written in a letter during the Battle of Balaclava, is another case in point, because time did not permit his addressee to respond.

For these reasons, I do not subscribe to the new orthodoxy. I will argue that "literary reference" is still only very partially understood and that, in fact, research in this area has become increasingly interesting in recent years. Readers curious about the computational modelling of referential collaboration are invited to read [Heeman and Hirst, 1995], [Engonopoulos et al., 2013], [Garoufi and Koller, 2014], and [Fang et al., 2014]; they will find there much that is of interest; moreover, they will find that all the issues discussed in the present book remain relevant when collaboration is taken into account.

Despite our focus on "literary reference", we will discuss the notion of *salience*, because it affects virtually all of our REs. Broadly speaking, an entity is salient for a person to the extent that it attracts this person's attention. In many cases, speaker and hearer attend to the same things, but sometimes they do not, as in the notorious Battle of Balaclava (chapter 1), where Lord Raglan and Lord Nolan look at a battle field from different vantage points, so different things were salient for the two of them. The role of salience in REG is discussed in sections 4.9 and 9.7.

ID	SPECIES	ORIGIN	WEIGHT	INJURIES	LOCATION
a	*lion*	*Kenya*	$102kg$	*paws, teeth*	**cage**
b	*lion*	*China*	$100kg$	*paws*	*?*
c	*tiger*	*China*	$310kg$	*back*	*?*

Table 1.3
Shared knowledge (expanded). New information is shown in boldface.

1.7 A Perspective on Reference: Information Sharing

Theoreticians have long debated the logical analysis of definite descriptions. In later chapters, I shall argue that most computational work on reference can be aligned with the insights of theoreticians who, in the wake of Peter Strawson, regard sentences that contain failed references as being neither true nor false [Strawson, 1950] (see section 2.4 for details). These theoreticians regard "The so-and-so is P" as consisting of two different parts, which contribute to communication in different ways. Implicit in this account is a distinction between information *presupposed* (i.e., shared or given) and information *asserted* (i.e., presumed new to the hearer, e.g., [Levinson, 1983], chapter 4). This distinction is fundamental to much theorizing about language but it is often ignored by computational linguists; it will play an important part in this book.

Our zoo may help to convey the idea. Suppose I have information about an animal a, for example, the fact that it is in the cage. If I want to communicate this information to you, and if our shared information is represented in the Knowledge Base depicted in Table 1.1, then I can make clear what animal I have in mind by saying, for example, "the Kenyan lion", making use of our shared knowledge. Suppose I choose this RE, then after my full utterance, "The Kenyan lion is in the cage", my *privileged* information has shrunk a little but our *shared* information has increased because the fact that a is in the cage is now a part of it. The result is an expansion of Table 1.1 that can be represented as Table 1.3.

We shall see in section 3.1 that it takes more for information to be shared than for this information to be known by all the people involved: information only counts as shared (also known as "common knowledge", or in "common ground") if all the people involved know that it is shared. A dramatic example is depicted in Figure 1.3, which shows a run on the banks, caused by the fact

Figure 1.3
A dramatic example of Information Sharing: a newspaper article causes a run on the banks.

that a newspaper report ("banks go bust!") becomes shared information. Crucially, a run on the banks does not result if every reader believes himself to be the only one to know that the banks are about to go bust: we will only run if we believe that others possess, or will soon possess, the same information.

Reference relies on information *shared* between speaker and hearer. This has interesting consequences. For example, if the speaker and hearer have different beliefs about a tiger's country of origin, then its country of origin is not in the Knowledge Base; only by "backing off" to a level at which the information is shared (e.g., the fact that the animal comes from Africa) can the origin of the tiger contribute to fixing the identity of the referent.

The interplay between shared (also called *given*) and privileged (also *new*) information, which Rodger Kibble and I once called *Information Sharing* [van Deemter and Kibble, 2002] for want of a more generally accepted term, is crucial to communication. In information sharing, information shared between speaker and hearer is exploited to enable the speaker to pass on her privileged information into the store of information that she shares with the hearer, as when she says "The Kenyan lion *is in the cage*".

Distinctions between given and new information come in different varieties (topic vs. comment; theme vs. rheme; topic vs. focus [Hajičová et al., 1998]; see also section 2.4 of the present book). It plays a role in a large variety of linguistic structures, involving not just REs but also factive verbs, cleft sentences, and counterfactual constructions. The distinction is also important to speech synthesis because new information tends to be marked by pitch accents that make the item in question stand out in the perception of the hearer (e.g., [Pierrehumbert and Hirschberg, 1990], [Gibbon et al., 2000]).

Researchers from a number of disciplines have contributed to our understanding of Information Sharing. The philosophers Stalnaker and Lewis, for example, offered a formal explication [Stalnaker, 1973], [Stalnaker, 1978], [Lewis, 1979], [Beaver, 1997]. David Lewis asked attention for situations in which a RE can add information, which may then be "accommodated" by the hearer; for example, I might say "My wife will be waiting" to someone who did not know that I am married (cf., section 7.4). In other cases, comprehension requires *deduction*: suppose we modify the example of Table 1.3 by assuming that the hearer does not know whether c is a tiger. Now the fact that is c is a tiger is not shared, yet if the speaker says "The tiger is in the cage", the hearer can combine the RE with the information in the knowledge base and understand which of the three animals is the referent. Such situations, in which accommodation or deduction is needed to make sense of an RE, will not be discussed in this book because they are, for the time being, beyond the state of the art of computational models of referring.

1.8 Summary of the Chapter

I have defined the scope of the book and the main assumptions behind it. We have seen that reference is central to communication, and illustrative of one of the key mechanisms underlying all communication, namely, *information sharing* (section 1.7). The following points are worth stressing:

- The book concentrates on the *production* of referring expressions; this will force us to ask why certain expressive choices are – or should be – made. The inverse problem, of understanding (i.e., interpreting) REs, will only be discussed where it is relevant for production; after all, by and large, speakers attempt to produce REs that are clear for hearers. [Section 1.3]
- We focus on issues specific to reference. This means that determining the semantic content of an RE will take precedence over finding the words

that express this content, because the latter happens in much the same way throughout language production. [Section 1.4]

- In much of the book, REs will be studied individually, in isolation from any linguistic context. This choice for "one-shot" REs will allow us to focus on some of the core mechanisms of reference. [Section 1.6]

2 Theories of Reference

Reference has featured in debates about language at least since Plato's examination of the concept of knowledge[1] and became an area of lively debate around 1900, when Gottlob Frege and Bertrand Russell started to use Formal Logic to shed light on the meaning of natural language.

This chapter will summarize some theoretical debates in this area. We will not cover all of these (cf., the full-length treatments of [Hawkins, 1978], the famous [Kripke, 1980]; [Bach, 1987], [Neale, 1990], [Recanati, 1993], [Reimer and Bezuidenhout, 2004], [Elbourne, 2005], [Abbott, 2010], [Kabasenche et al., 2012], and [Hawthorne and Manley, 2012]) but focus on what is most relevant for reference production instead. Later chapters will often ask whether the challenges posed by theoreticians have been adequately tackled by computational models.

The plan of this chapter is as follows. After discussing what may be meant by the term "referring expression" (section 2.1), we shall reflect on the key notions of unique identification (section 2.2), denotation, and connotation (section 2.3). We shall discuss the famous debate between Russell and Strawson on the proper analysis of definite descriptions (section 2.4), after which we shall reflect briefly on substitutivity in intensional contexts (section 2.5). Attributive descriptions and misdescriptions will be introduced (section 2.6), after which we turn to proper names (section 2.7) – a topic that looms large in philosophical studies of reference – and the Gricean Maxims (section 2.8).

2.1 What Makes a Referring Expression?

The difficulty of pinning down the notion of an RE can be illustrated aptly using an episode when researchers in Information Extraction proposed a scheme for allowing annotators of language corpora to say which expressions refer to the same entity [Hirschmann and Chinchor, 1997]. Annotations of this kind are necessary if you want to create a "gold standard" for testing

1 Plato's *Theaetetus* dialogue asks what it means to be able to *"mention some mark which differentiates the object in question from everything else. (...) Take the sun for example, if you like: I think you'd be on safe ground with the idea that it is the brightest of the heavenly bodies which travel around the earth* (sic!) *(...) if you get hold of what uniquely differentiates something from everything else, you will arguably get a rational account of just that thing; but if the feature you get hold of is shared, your account will be concerned with however many things share this feature"* ([Waterfield, 1987], section 208b). As we shall see in the next chapter, Plato's focus on uniquely differentiating properties echoes contemporary concerns.

coreference resolution programs (see section 1.2). Drawing up a good annotation scheme requires a clear understanding of the phenomena annotated. As it happens, the Hirschmann-Chinchor scheme, which came with a written explanation known as a task definition, was riddled with inconsistencies [van Deemter and Kibble, 2000]. For instance, the task definition encouraged annotators to disregard the role of time in the interpretation of REs. For example, in the sentence

> Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents,

annotators were asked to mark *sales director of Sudsy Soaps* as coreferential with both *Henry Higgins* and *president of Dreamy Detergents* (cf., [Cristea et al., 1999]). But because coreference is an equivalence relation (reflexive, symmetrical, and transitive) this implies that the sales director of Sudsy Soaps and the president of Dreamy Detergents must be the same person. Clearly this cannot be right, and a reasonable conclusion is that, of the three NPs, only *Henry Higgins* refers [van Deemter and Kibble, 2000]. The other two NPs, which are predicatively used, do not.

How can we tell whether an NP is an RE? Linguists have devised syntactic tests to distinguish between different kinds of NPs, but these appear to be of limited use. Milsark, for example, stipulated that NPs are *weak* if they can fill the NP slot in sentences of the form "There exist(s) NP" and *strong* if they cannot [Milsark, 1977]. Might the notion of a strong NP help us to define the notion of an RE? Clearly, definite descriptions are strong, and so are NPs of the form "every so-and-so". On the other hand, Milsark's criterion also makes NPs of the form "most so-and-so's" strong, yet, by themselves, they do not refer to any specific set of so-and-so's. In fact, one should be skeptical of any purely syntactic criterion, given that one and the same NP can be used both referentially and nonreferentially. Consider, for example,

> One German bank is on the verge of collapse.

This may be a statement about the number of German banks that are at risk, but it could also be about a specific bank. A purely syntactic test would not be able to tell the difference between these two different interpretations.

Approaches oriented towards Formal Logic have more to offer, but tend to fall short as well. Perhaps most notably, Barwise and Cooper proposed a perspective in terms of the formal theory of Generalized Quantifiers ([Barwise and Cooper, 1981]; see [Peters and Westerstahl, 2006], section 4.6,

for discussion). Informally, the theory of Generalized Quantifiers lets an NP denote a set of sets. For example, the NP "two people" denotes the set of all sets that contain (any) two people, the NP "the red circles" denotes the set of all sets that contain all the red circles, and so on. Using this perspective, an NP is *semantically definite* if it can be characterized by the algebraic concept of a nontrivial *principal filter*. A set of sets is a nontrivial principal filter if it can be described as $\{X : Y \subseteq X\}$ (i.e., the set of all supersets of Y), for some nonempty set of entities Y.

To see the implications of this definition, consider "President Obama". This noun phrase is semantically definite because, in all domains in which it denotes, it denotes the set of all subsets of the domain that contain President Obama, which can be written (verbosely) as $\{X : \{\text{Obama}\} \subseteq X\}$, which is a principal filter. Similarly, "the red circles" is definite, because when it denotes, it denotes the principal filter $\{X : C \subseteq X\}$, where C is the set of all red circles in that situation. An NP like "two people", however, is not semantically definite, because the set of all sets that contain (any!) two people is not a principal filter. The same is true for "most people", as one can easily see.[2]

Formal analyses can be intellectually pleasing yet difficult to apply to concrete cases. In the example (above) of the expression "one German bank", for instance, consider a situation in which the speaker has a specific bank in mind. Now some might argue that the NP is semantically definite, hence referential; after all, it denotes an entity b that the speaker would be able to describe uniquely, hence the NP denotes the set of all sets that contain b as an element, which can be written as $\{X : \{\text{b}\} \subseteq X\}$. On the other hand, the *hearer* would not be able to identify b on the basis of the utterance above, therefore others might argue that the NP does not refer. There is no unanimity on these matters, although it seems that in recent years most theoreticians have taken the former position [Dekker, 1998], [Schwartzschild, 2002], [Breheny, 2008] [Hawthorne and Manley, 2012].

Despite these difficulties, one might regard semantic definiteness as a *necessary* condition for being an RE. For instance, NPs like "no-one", "most people", "fewer than five people", and "at least five people" are not semantically definite; and indeed, it would be difficult to say which precise individuals they refer to. Similarly, bound anaphors are not semantically definite; and indeed, it

2 A different perspective is offered in [Kamp and Reyle, 1993], chapter 4, where the subject of "Most linguists use a parser" refers to the set of those linguists *who use a parser*. Their perspective captures what entities are introduced into a discourse by a sentence as a whole and become available as antecedents for anaphora. See also [Constant, 2012].

would be unusual to say that "it", in "Every TV network reported its profits", refers. Whether semantic definiteness should be seen as a *sufficient* condition for being an RE seems more doubtful; universally quantified NPs, and generic NPs, are semantically definite, for example, but it seems doubtful that they refer in the everyday sense of the word (*pace* [Shaw and McKeown, 2000]).

Ultimately, however, quibbling over definitions is not productive. After all, a definition cannot be true or false – it can only be more or less useful (and more or less in accordance with existing usage). In chapter 4, I will offer a definition of *the classic* REG *task* (section 4.3), and this purely stipulative definition will guide our discussions for a while. In Part IV of the book, we shall open up to a much wider range of "reference" tasks, each of which will be introduced not by means of a formal definition but by means of a concrete scenario of communication between people.

2.2 Knowing What Something Is

Searle wrote, "Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular I shall call a referring expression", but what does it mean to identify something? Entire treatises have been devoted to this question (e.g., [Strawson, 1959]). To see what's at stake, let's consider an example from [Aloni, 2002]. An Ace of Spades and an Ace of Hearts are laid out in front of you, one to the left of the other. The cards are lying face down, so you do not know which one is which:

♠ ♡ , or

♡ ♠

You're told that one of these cards is worth a large sum of money, leaving you to guess which one. There are two types of knowing which card wins you the money: one can know this in terms of the position of the cards or in terms of their suit. Aloni calls these types of knowing *conceptual perspectives*. For example, you may know that *the Ace of Spades* wins, or you may know that *the card on the left* wins. Different situations favour different conceptual perspectives. In the situation at hand, it's no use knowing that *the Ace of Spades* wins, but it's extremely useful to know that *the card on the left* wins. If, however, you asked a friend, "Which card wins?", then an answer that used the perspective of suits would be uninformative. If, instead of asking which card wins, you asked, "Which card is that?" (pointing to the card on the left), then a third

conceptual perspective would be preferred, as in "It's the winning card". A speaker's choice between these perspectives can have important consequences.

Conceptual perspectives are not restricted to games. In fact, the first systematic study of this topic, by the philosophers Boër and Lycan, focussed on the identification of people in real-life settings [Boër and Lycan, 1986]. Suppose I'm interested in knowing who has won the presidential elections that took place, in a certain country that I do not know much about, yesterday. Which of the following situations guarantees that I know who has won?

- I know the person's full proper name and address. I know nothing else about the person, not even their gender, age, and position on the political spectrum.
- I've seen the person in an elevator and remember the person's appearance. I know nothing else about the person, including the person's name.
- I know that this person was the leader of a protest movement that swept aside the previous government. I know nothing else about the person.

The information in each of these situations may be so specific that it can only apply to one person. Yet, it is unclear which of the above lets me know who won. Examples of this kind show that "identifying the referent" isn't always the same thing: the referent needs to be identified using information of "the right kind", and no single kind is always right.

Boër and Lycan thought that in choosing a conceptual perspective, the *purpose* of the identification matters especially ([Boër and Lycan, 1986], pp. 33-37). This purpose suggests a particular *category* (e.g., the address or the portrait or the political party of the referent) and a particular *format* (e.g., in the case of an address, the answer should provide a street name and number).

The computational literature contains echoes of these issues. For instance, [Appelt, 1985b] proposed that to refer successfully is to produce a description $\mathcal{D}$ such that there exists another description $\mathcal{D}'$ such that the hearer knows that $D = D'$ and D' is, *prima facie* identifiable. The author stipulated that such terms "meet certain syntactic criteria for being the 'right kind' of term" (echoing Boër and Lycan's "format"). For example, suppose I am the postman, and I ask you to whom I should deliver a letter. You respond, "deliver it to the person whose name is on this envelope", while the envelope has both the name ("Mr. so-and-so") and the address "41 Malvedere Crescent, Aberdeen". Then $\mathcal{D}$ = "the person whose name is on this envelope" and $\mathcal{D}'$ = "the person named Mr. so-and-so, living at 41 Malvedere Crescent, Aberdeen, UK". The

reference is successful provided the combination of the name and the address meets the syntactic criteria set by the post office.

An ability to always choose the right perspective might be seen as the Holy Grail of REG, and as such it is a topic that runs through most chapters of this book. For example, in sections 3.4 and 8.7, we shall see how the notion of an *attribute*, as a conglomerate of related features, implements the notion of a conceptual perspective; in chapter 13, we shall explore how questions of the form "Who is ...?" may be answered by a computer program. Before we delve deeper into these issues, it will be useful to see how philosophers of language have analysed the meaning of REs and the sentences containing them.

2.3 Denotation and Connotation

The 19th-century economist and philosopher John Stuart Mill appears to have been the first to distinguish between the denotation and the connotation of an expression [Mill, 1843]. The Millian *denotation* of a word like "animal" is simply the set of all animals. Its *connotation* is the set of those properties that something needs to have in order to count as an animal: locomotion, procreation, metabolism, and so on. Similarly, the denotation of "even number" is the set of numbers $\{0, 2, 4, ..\}$; the connotation is "being divisable by 2". By a slight extension of Mill's usage, we also speak of the denotation of a property.

Mill's term "denotation" is still in use; Rudolf Carnap's word "extension" is often used as a synonym [Carnap, 1947]. In fact, most REG algorithms can be described entirely in Millian terms: they search for a combination of properties whose denotations are such that, when these denotations are intersected, the resulting set contains the referent and nothing else. In discussing REG, we shall frequently use the notation $[\![x]\!]$ to talk about the denotation of x. We shall do so when x is a word or a linguistic expression but also when x is a semantic construct that is used to represent a word or an expression. Usually, we take $[\![x]\!]$ to be limited to some particular domain of discourse. For example, $[\![zebra]\!]$ might stand for the set of zebras in a particular zoo. If the intersection $[\![zebra]\!] \cap [\![pregnant]\!]$ happens to be a singleton set, then the combination of the properties "being a zebra" and "being pregnant" is suitable for building an RE. This RE would refer to the set containing only the one pregnant zebra and, by an obvious extension, to this zebra itself.

At the end of the 19th century, the philosopher and logician Gottlob Frege wrote extensively about the way in which the meaning of a sentence depends

on the meanings of its syntactic parts [Frege, 1960], a perspective that has come to be associated with the term *compositional semantics*. Following in the footsteps of Mill, Frege distinguished between the *Sinn* (henceforth: sense; similar to Mill's connotation) and the *Bedeutung* (henceforth: reference; similar to Mill's denotation) of an RE. The latter is simply the referent, whereas the former can be described loosely as the way in which the referent is presented.

Frege's examples include expressions like *the morning star* and *the evening star*: both refer to the planet Venus, but in different ways, using different definitions, one might say. Frege observed that, usually, the truth of a sentence does not depend on the sense of the REs in it but only on their reference. For example, if it is true that "the morning star is a solar planet composed of rock", then the sentence "the evening star is a solar planet composed of rock" must also be true. The context "x is a solar planet made of rock" is normal in this regard. Crucially, however, some contexts – which came to be known as *intensional* contexts – are different in that their truth depends not just on the reference of x but on its sense. An example is the context created by "Homer Simpson knows that x is a solar planet composed of rock". The sentence that results from letting "the morning star" take the place of x might be true, whereas the result of letting "the evening star" take its place might be false.

To vary on this theme, consider the context "Homer Simpson knows that x is dangerous". This may be true for $x =$ "the rocket launching button" yet false for "the red button at the top of the console", even if the rocket-launching button *is* the red button at the top of the console. Contexts like this, whose truth value depends on the sense (not just the reference) of the REs in them, are called *intensional*. Later theorists speak of *hyper*-intensional contexts in situations where even the sense of the REs does not suffice to determine the truth value of the utterance.

Frege was able to offer an elegant account of the compositional semantics of sentences that contain intensional contexts, as well as ones that do not. Later, more detailed work by Rudolph Carnap, Richard Montague, and others can be seen as making Frege's ideas precise. Many of these ideas still stand, though numerous refinements and additions have been proposed. Some of the most important refinements are associated with the names of Russell and Strawson.

2.4 The Russell-Strawson Debate

In a famous study, Bertrand Russell applied Predicate Logic to the analysis of English sentences [Russell, 1905]. In order to get a clear perspective on the role of definite descriptions, he asked how one should analyse sentences that include a description where "the so-and-so" fails to pick out one unique individual. Examples include sentences such as "The King of France is bald", "The King of France is not bald", and so on. Are such sentences false, or does there arise a "truth value gap" (i.e., a situation in which the sentence is neither true nor false)? Russell offered an account of such sentences that does not allow for truth value gaps. To him, "The King of France is bald" *asserts* that there exists a unique King of France: with K as short for "is a King of France", this can be expressed by the Predicate-Logical formula

$$\exists x(K(x) \wedge \forall y(K(y) \rightarrow y = x))$$

("there exists exactly one King of France"). The sentence "The King of France is bald" can now be rendered as follows, with B as short for "bald":

$$\exists x(K(x) \wedge \forall y(K(y) \rightarrow y = x)) \ \wedge \ \forall z(K(z) \rightarrow B(z))$$

("there exists exactly one King of France, and all Kings of France are bald"). Russell regarded the pattern displayed here as so important that he introduced a specific logical symbol, the iota (ι) operator, to abbreviate it. Just like the meaning of "a King of France" can be rendered directly using the existential quantifier, writing $\exists x K(x)$, the meaning of "the (unique) King of France" could now be rendered directly using the iota operator, writing $\iota x K(x)$. The meaning of the sentence as a whole can be rendered as $[\iota x K(x)]B(\iota x K(x))$, in which the description occurs twice.[3]

Russell's analysis allows him to shed light on more complex sentences, and the iota operator allows this analysis to be stated succinctly. For example, Russell analyses "The King of France is *not* bald" as ambiguous between a true and a false sentence: it can either mean $[\iota x K(x)]\neg B(\iota x K(x))$, which is false (because it asserts the unique existence of a King of France), or $\neg([\iota x K(x)]B(\iota x K(x)))$, which is true (because the assertion as a whole is

3 Russell used a special logic notation that is no longer in use. As is common practice, we explain his ideas by mixing his ι operator with modern notations.

negated). The legitimacy of this interpretation becomes clear when we consider "The King of France is not bald, because France is a republic". Similarly, using $\diamond$ as the possibility operator of Modal Logic, "The King of France *may* be bald" is ambiguous between $[\iota x K(x)] \diamond B(\iota x K(x))$ (which asserts the existence of the King of France) and $\diamond([\iota x K(x)] B(\iota x K(x)))$ (which only asserts the possibility of his existence), both of which are genuine interpretations of the sentence. The fact that Russell was able to offer attractive analyses of problematic sentences like these is one of the reasons why it is justly famous.

Despite its virtues, Russell's account is sometimes regarded as linguistically crude, because it fails to distinguish between assertion and presupposition. In Peter Strawson's view, "the King of France" does not *assert* the existence of a king: it *presupposes* it [Strawson, 1950]. Strawson therefore regarded "The so-and-so is P" as consisting of two parts, which contribute to communication in different ways: the first part, "The so-and-so", nods towards a store of information assumed shared between speaker and hearer, whereas the second part, "is P", adds to this information by making an assertion.

As was noted in section 1.7, the computational literature has generally accepted the Strawsonian viewpoint as crucial to the task of computationally generating REs. This same division of labour occurs in a variety of linguistic constructs, loosely associated with the term presupposition (e.g., [Levinson, 1983], chapter 4; see also our section 1.7). They include factive verbs (like "regret" and "realize"), change-of-state verbs (like "stop"), and so on. For example, consider the sentence "Alice stopped nagging her husband". A broadly Russellian analysis would have this mean that Alice used to nag her husband and now she does not. Strawson would say that this misses a nuance, namely, that the speaker doesn't so much assert the nagging but presupposes it. The same is true for definite descriptions such as "The King of France is bald": such descriptions treat the existence of a unique King of France as a given. Other objections have been raised against the Russellian analysis,[4] but the inability of the Russellian analysis to do justice to the distinction between assertion and presupposition stands out as the most important one.

In recent years, researchers have argued that it may not be necessary to choose between Russell's and Strawson's accounts [Donnellan, 1966],

4 One issue is that, taken literally, the Russellian formula $[\iota x \phi(x)] \psi(\iota x \phi x)$ requires that there is only one x with the property $\phi(x)$ in the entire model (i.e., without restricting to x that are contextually salient), which is often implausible. This issue is no longer regarded as a deep problem for the Russellian analysis [Reimer and Bezuidenhout, 2004], [Neale, 1990] [Dekker, 1998], however. For a discussion in the context of REG, see section 9.7.

[Neale, 1990], [Dekker, 1998]. Some have argued that each account matches a particular *use* of definite descriptions (see section 2.6). It is this position that we shall take. In Part IV, for example, we shall encounter expressions that, while referring in some situations, play a Russellian role in others.

Where do these discussions leave *in*definite descriptions? One answer is that indefinite descriptions are very different from definite ones, because they are existential quantifiers, asserting that there exists at least one object that has a certain property. Consider a sentence discussed by [Dekker, 1998]: "A painting is missing from the museum." According to the standard account, this expresses the proposition

$$\exists x(P(x) \wedge M(x)),$$

where P stands for Painting and M for Missing. On closer reflection, however, "A painting" may be uttered in two different types of situation. In a scenario of type **A**, it is uttered by a policeman who checks the walls of the museum and notices an empty spot on the wall. This is a nonspecific use of the indefinite if the speaker does not know which painting is missing. This usage is captured adequately by the account above. But consider a scenario of type **B**, where the sentence is uttered by an art lover who has discovered that his beloved *Who's afraid of Red, Yellow and Blue II* is no longer on the walls of the hall that it once adorned. This is known as a *specific* use of an indefinite, because the speaker has a concrete painting in mind. Expressions of this kind may be thought of as REs without the uniqueness presupposition.

Others have called the existence of specifically used indefinites into question, arguing that the difference between the two scenarios mentioned above is not as dramatic as it seems [Dekker, 1998]. After all, bearing in mind the conceptual *perspectives* of section 2.2, there is a perspective that does let the policeman of scenario **A** identify the painting: it is *the painting that hung at location* x, where x is a specific space on the wall (e.g., where the wallpaper stands out as slightly lighter than elsewhere). Surely this is a unique reference to the painting.

Theoreticians today believe that definite and indefinite NPs have much in common, so they are often discussed together, as different sides to the same coin [Hawkins, 1978], [Neale, 1990], [Reimer and Bezuidenhout, 2004]. To use the same nautical analogy that we made at the start of the book, both types of NPs can "anchor" words to things. Some authors have even denied that there exists a systematic difference between the two types of descriptions.

This school of thought denies that definite descriptions express uniqueness, citing examples such as

John went to *the* dentist.
John was hit in *the* eye.
Let's go to *the* pub.
She is *my* student.

as evidence, because they contain a definite description that does not express uniqueness (see [Ludlow and Segal, 2004]; also [Ludlow and Neale, 1991]). Ludlow and colleagues argue that uniqueness is only imparted by our background knowledge. For example, the fact that "the King of France" is interpreted as describing a unique king of France does not stem from the definite article – which many languages do not express – but from our background knowledge about kings (i.e., the fact that there tends to be only one of them in any given country at any given time).

I will not pass judgment on these matters. Instead, we shall focus on algorithms that single out a referent by finding semantic properties that allow a recipient to single out the intended referent; we shall have much less to say about the way in which these properties are to be put into words, including the choice between the definite/indefinite article. In fact, we shall see that the distinction between the two kinds of descriptions can become even more blurred when it is difficult to determine whether a given description singles out a unique referent or not (chapter 15).

2.5 Intensional Contexts

A venerable strand of theoretical research studies the logical properties of modal, temporal, and epistemic contexts (e.g., [Linsky, 1971], [Neale, 1990], [Groenendijk et al., 1996]) and a significant amount of this work focusses on descriptions. A typical question is what other descriptions one can substitute for a given description without risking a change of truth value (i.e., *salva veritate*). One might think that if two descriptions have the same denotation (i.e., they are co-extensive) they can always be substituted for each other *salva veritate*; in the words of Stephen Neale,

Principle of Substitutivity: If (i) $a = b$ is a true identity statement, (ii) α is a true sentence containing one or more occurrences of a, and (iii) β is the

result of replacing at least one occurrence of a in α by b, then (iv) β is also true. ([Neale, 1990], section 4.3.)

This Principle does not always hold, however. Consider the following sentence,

Next year, the round button will be located at the top of the console,

uttered in the cockpit of an airplane, which is full of buttons and dials; the red button is round, and there is only one red button on the console, and only one round one. Consequently, the red button = the round button. Yet if we substitute "the red button" for the subject of the sentence, then the truth value of the sentence may change. Contexts like this, in which the Principle of Substitutivity fails to hold, are called *intensional* (also referentially opaque), in contrast to *extensional* contexts, where the Principle does hold. They include:

1. *Contexts created by temporal and modal operators:* "Next year, the red button will be at the top of the console." "The button must be red".
2. *Contexts created by epistemic contexts:* "The manager believes that button is at the top of the console". "Everyone knows that the button is at the top".

Some substitutions into these contexts do preserve truth, but these require equalities stronger than in clause (i) of the Principle of Substitutivity. Suppose, for instance, that round buttons are also known as dials, so "round button" and "dial" are synonymous. Suppose we substitute the latter for the former; this substitution preserves the truth values in intensional contexts.

Does this work have consequences for computational models?

First, as the story of this book unfolds, it will become clear that generation in intensional contexts is beyond the reach of existing REG algorithms. This is understandable, given that in these contexts, the notion of reference is problematic: if the descriptions in them refer at all, then it is not to an object in the real world but to an object in some other (future, imagined, etc.) world. Theories have interesting things to say about these contexts, but they do not yet offer the detail and precision required by computational REG models. This is no coincidence: as we shall see, computational models tend to lag behind pure theory, with theories exploring issues long before they are addressed by means of algorithms and computer programs.

Second, intensional contexts highlight the importance of choosing an appropriate conceptual perspective (section 2.2). Consider the choice that a speaker might face between two descriptions in a simple extensional context:

The round button / *The rocket-launching button* is dangerous.

Suppose the two REs are co-extensive, so the truth of the sentence does not depend on the choice between them. Yet they cannot be used interchangeably. For example, in response to the question, "Are there any dangerous buttons in the cockpit?", the latter choice (which indicates the function of the button) will tend to be more felicitous than the former. On the other hand, synomyms can usually replace each other: for example, any context that makes "the round button" felicitous makes "the dial" felicitous as well, and conversely. More generally, it seems reasonable to hypothesize that two descriptions can be used interchangeably "*salva felicitate*" in extensional contexts only if they are interchangeable *salva veritate* in intensional contexts. In this roundabout way, intensional contexts might be relevant to present-day REG after all.

To sum up, problems usually associated with intensional contexts plague the generation of REs in extensional contexts as well. The risk, in the extensional case, is not that sentences may be generated that are *false*. The risk is that sentences may be generated that are uninformative, infelicitous, or clumsy.

2.6 Attributive Descriptions and Misdescriptions

It is time to introduce a distinction between two ways in which descriptions of the form "the so-and-so" can be used. The precise demarcation between them is a matter of debate (see [Abbott, 2010], section 6.3, for discussion), but there are clear cases on either side of the distinction.

Consider the sentence "The green button (is at the top of the console)", as spoken to someone who can see the button. The subject NP is used *referentially*. The NP doesn't contribute to the meaning of the utterance, except by contributing a referent (see [Recanati, 1993] for extensive discussion of this idea of *direct* reference). A simple test confirms that my description was referential: if I were told that my perception of the colour of the button is mistaken (e.g., because of temporarily strange lighting conditions in my room), then I'd want to revise my description, ascribing the correct colour (e.g., blue) to the button. Referential descriptions are captured well by Strawson's account (section 2.4), and it is them that most REG algorithms try to produce.

But definite descriptions can also be used in another type of situation. Suppose I say "The rocket-launching button is dangerous", speaking from a general knowledge of rocket launchers. The gist of my utterance may be conveyed as saying that whatever button is used for launching rockets, it is dangerous.

In situations of this kind, the NP is used *attributively*. This time the test that was used above confirms that my description was not referential: suppose you told me that the button on the panel in front of me, which I thought to be the rocket launching button, actually serves a different purpose. This new information would not cause me to want to revise my original description, because it remains correct. Attributive descriptions are captured well by Russell's account (section 2.4), which makes the descriptive content an essential part of the assertion made by the sentence containing it.

Attributive descriptions were discussed in [Donnellan, 1966], [Grice, 1969] and elsewhere. To use an example from Donnellan, if we say "The murderer of Smith is insane", this can be used referentially, to denote a concrete individual believed to have murdered Smith, but it can also be used attributively, even before anyone knows who has murdered Smith. The circumstances of the murder may have led the speaker to deduce that the murderer, whoever she is, must have been insane.

Attributive descriptions require a deep understanding of the domain of discourse and an ability to reason. Computationally modelling the production of such descriptions is challenging but potentially rewarding as well. After all, attributive descriptions embody an even more interesting aspect of language than referential ones: the latter can be replaced by pointing, so the use of language could be argued to be, in some sense, incidental. Attributive descriptions, by contrast, are essentially linguistic and closely tied up with the ability (which is often thought to be specifically human) to think and talk about things that are not "here and now". In section 10.6, we will indicate how attributive descriptions can be modelled.

In another example from [Donnellan, 1966], a speaker talks about a man at a cocktail party who is holding a glass, saying, "The man with the Martini is the murderer of Smith". Suppose the name of the man she intends to refer to is Jones, but Jones drinks wine. No one at the party drinks a Martini; crucially, however, Jones did murder Smith. Clearly, something has gone wrong, but has a falsehood been uttered? Many researchers believe that the sentence is *true* even though it misdescribes Jones. The Russellian analysis disagrees, because it understands the sentence as stating, among other things, that exactly one person at the party drinks a Martini, which is false. The Strawsonian analysis has less difficulty with the situation, allowing "The man with the Martini" to refer to Jones, even though the manner in which it does so (i.e., by presupposing that

Jones was drinking a Martini) is incorrect. According to the Strawsonian analysis, the assertion contained in the utterance (although not the presuposition contained in it) is true.

Another problematic class of descriptions involves metonymy, as in the following example from [Nunberg, 1978]: a waitress says this to a colleague, referring to a customer whose proper name she doesn't know but who is waiting for his sandwich to arrive: "The ham sandwich is getting restless". Although the intended referent is not himself a ham sandwich, he has ordered one. This fact is used by the waitress to produce a succinct description of the man in a situation where a description that is literally correct would have tended to require more words (e.g., "The man who has ordered a ham sandwich", "The man who is sitting on his own, near the exit").

I have little to say about misdescriptions, but there is scope for computational models here. A good model would embody a theory of why and when a misdescription makes a natural and effective RE. For example, the question is whether Donnellan's utterance can ever be the speaker's best option. Perhaps it can: if all that is known is the look of the drink, then *the man with the Martini* may be the most succinct RE that identifies the drink in a useful manner. An analysis along these lines may have something to add to existing theories (e.g., by shedding light on the difference between semantic and pragmatic reference [Kripke, 1977], [Neale, 1990], [Devitt, 2004]).

Metonymic descriptions may yield to the same method. For although Nunberg's "the ham sandwich" is not literally correct, it may be the speaker's best way to identify the referent. The challenge for computational models, in all these cases, is to exploit nonliteral meaning, deviating from the literal truth where this makes communication more effective.

2.7 Proper Names

A remarkable proportion of philosophy research on reference concerns *proper names*.[5] The most widely discussed question can be worded in Fregean terms: do proper names have a *sense*? Mill (cf., section 2.3) believed that proper names do not have a sense at all (using his terminology: they do not have connotation), but only a reference (Mill: they only have denotation) [Mill, 1843]. This position was challenged by Frege, who provided examples such as

5 See e.g., [Cumming, 2013], from whose exposition this section will borrow freely.

Homer believed that the Morning Star was the Evening Star.
Homer believed that the Morning Star was the Morning Star.

The second of these sentences can only be true, of course, but the first is false. If "the Morning Star" and "the Evening Star" are proper names, and if the *sense* of a proper name equals its referent (as Mill claimed), then both sentences should have had the same truth value given that both names denote the planet Venus. The argument does not hinge on the choice of these particular REs: if we replace the first one by *Phosphorus* (another name for the Morning Star) and the second by *Hesperus* (another name for the Evening Star), then Frege's argument still goes through. Most later theoreticians have largely accepted Frege's criticism of Mill's position, although they have differed in terms of their preferred solutions.

A once-popular solution, which still has considerable appeal, is to regard proper names as implicit descriptions. "The Evening Star", for example, might be shorthand for "The first star to light up in the evening". Similarly, the proper name "Aristotle" might abbreviate something like "the last great philosopher of antiquity". In this way, proper names inherit their sense from the description that they abbreviate, thereby evading the problem noted by Frege. If it is countered that these descriptions are too arbitrary – why, for example, does "Aristotle" not mean "Plato's most famous pupil" instead? – then one might resort to regarding a name x for a person as abbreviating the description "the person named x".

The theory that proper names are implicit descriptions came under attack when Saul Kripke's famous *Naming and Necessity* provided a range of counterexamples and proposed an alternative account [Kripke, 1980]. His counterexamples include contrasting pairs like this:

Aristotle might not have been the last great philosopher of antiquity.
Aristotle might not have been Aristotle.

The first of these sentences states a perfectly plausible fact, because Aristotle's life could have taken a different course or because antiquity might have produced another great philosopher after him. The second sentence, however, is plainly false. It follows that "Aristotle" does not mean "the last great philosopher of antiquity".

Kripke believed that even simple sentences can be affected by these issues.In the Preface to [Kripke, 1980], for example, he hints at an imaginary written history whose author devotes a section to what would have happened had Aristotle

never become a philosopher, so Plato would have been the last great philosopher of antiquity. If the author wanted to elucidate this story with a picture of Aristotle, then surely he should have used a picture of "our" Aristotle, not one of Plato. Yet if "Aristotle" meant "the last great philosopher of antiquity", then a picture of Plato would have been more appropriate.

Counterfactuals led Kripke to observe that proper names have a constancy – they refer to the same individual in each possible situation – that descriptions lack: for this reason, he called them *rigid designators*. Constancy can be modelled elegantly in Modal Logic. Constancy also led Kripke to propose that proper names acquire their meaning through an act that can be likened to baptism. According to Kripke's proposal, the fact that you and I were absent when the name "Aristotle" came to be associated with Aristotle is irrelevant: if we use the name correctly, then we use it to refer to the person with whom it came to be associated.[6] Crucially, this relation between a name and the thing it denotes remains constant even in sentences that consider states of affairs that differ from the actual world, as in "Aristotle might not have been ...", or "If he had accepted Apartheid, Nelson Mandela would have been long forgotten".

The analysis of proper names is still a matter of debate (e.g., [van Langendonck, 2007]), to which various types of evidence have been brought to bear, including data from pathology [Semenza and Zettin, 1989] and neuroscience. I shall ignore these issues until demonstrating, in chapter 7, how proper names can be incorporated into computational REG.

2.8 The Gricean Maxims and Relevance Theory

We feel we understand a person's action if we can see that the action was chosen for a good reason. Accordingly, academic theories of human action often hinge on the idea that actions take place for a reason. One of the most celebrated theories of human communication, embodied in the Gricean Maxims, is of this type [Grice, 1975]. Grice proposed four Maxims, known by the names of Quality, Quantity, Relation, and Manner, which jointly implement the idea that communication is normally meant to be cooperative (Grice called this the Cooperative Principle), and which we render here in his own words.

6 An analogous theory was proposed by Hilary Putnam to account for the meaning of *natural kind* terms such as *water*, *aluminum*, etc., which can be regarded as referring as well [Putnam, 1975]. See Kripke [Kripke, 1980], lectures I and especially III, for a comparison with Putnam's views.

Quality. The Maxim of Quality requires two things:
1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Quantity. The Maxim of Quantity:
1. Make your contribution as informative as is required.
2. Do not make your contribution more informative than is required.

Relation. The Maxim of Relation contains just one requirement:
1. Be relevant.

Manner. The Maxim of Manner says:
1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

Grice notoriously did not formalize his Maxims, which are consequently open to different interpretations. Importantly, the Gricean Maxims come into their own when they are "honoured in the breach": they help speakers to convey information by violating them (or appearing to violate them). One of Grice's own examples is a situation in which someone says, "Mrs. Jones made some sounds which approximated the score of Home Sweet Home". By violating brevity (an aspect of the Maxim of Manner), the speaker conveys the information – known as a *conversational implicature* – that Mrs Jones wasn't very faithful to the original song. Another example involves a reference letter for an academic, which says, "The applicant has attended my lectures punctually and has a nice clear handwriting" and nothing else. By *not* saying how brilliant the candidate is, thereby appearing to breach the Cooperative Principle, the author of the reference letter makes it clear that the candidate should not be hired. Interestingly, once the recipient of the letter has "joined the dots", the Cooperative Principle has been restored again, because all the required information (i.e., that the application lacks the necessary academic qualities) has been conveyed. Exactly why the intended implicatures are derived (instead of e.g., that the recipient of the message knows the academic qualities of the candidate already) is a matter for debate.

One development from the Gricean Maxims has to be mentioned here, namely, Relevance Theory [Wilson and Sperber, 2004]. This theory emphasizes the importance of relevance, arguing that the Maxim of Relation, properly understood, can do the work of all the Maxims combined, and even more than that. The core idea is a notion of utility: speakers maximize the number of "relevant" inferences that can be drawn from an utterance:

> When is an input relevant? Intuitively, an input (a sight, a sound, an utterance, a memory) is relevant to an individual when it connects with background information he has available to yield conclusions that matter to him: say, by answering a question he had in mind, improving his knowledge on a certain topic, settling a doubt, confirming a suspicion, or correcting a mistaken impression [Wilson and Sperber, 2004].

Applying this idea to reference, [Wilson, 1991] gives the following example:

(a) I switched from linguistics to geography.
(b) The lectures were too/less boring.

Wilson argues that if the (b) sentence says "too boring", the intended referent of "the lectures" must be the linguistics lectures, but if it says "less boring", the referent has to be the geography lectures. This is difficult to understand without reasoning about the *utility* of the interpretations considered (which is to be maximized), and about the cognitive effort required by the reader (which needs to be minimized). The potential implications for NLG are obvious: if (b) is the utterance that allows the hearer to draw the relevant inferences with minimal cognitive effort, then presumably it is preferable over an alternative utterance such as "The linguistics lectures were too boring", because this is over-elaborate. Although these particular examples concern relatively complicated ("bridging") REs, we shall see in section 4.9 that similar considerations apply to simple REs.

2.9 Summary of the Chapter

The Philosophy of Language has produced a terminology – including the opposition between extension and intension, and between attributive and referential descriptions – that will stand us in good stead in later chapters, where psycholinguistic experiments and computational algorithms are discussed. Additionally, a number of lessons contained in this chapter are worth highlighting:

- The phrase "identify the referent" suggests a clarity that does not always exist [Appelt, 1985b], [Boër and Lycan, 1986], [Aloni, 2002], because identification can serve different purposes, and different purposes can involve different conceptual perspectives. [Section 2.2]
- The majority of work on computational REG relies on a distinction between given and new information (cf., section 1.7), thereby making Strawson's position, in his famous debate with Russell about the meaning of definite descriptions, concrete. [Section 2.4].
- Although definite and indefinite descriptions differ in some respects, there may be more that the two classes of descriptions have in common than there is that separates them. [Section 2.4].
- Intensional contexts teach us a lesson that has implications in extensional contexts as well, namely, that two REs are rarely identical in terms of the communicative situations in which they can be used. [Section 2.5]
- Attributive descriptions are occurrences of NPs whose descriptive content is more important than their referent; utterances that contain an attributive description are best seen as being about whoever, or whatever, meets the description (section 2.6). Although REG has rarely focussed on them, recent work (see chapter 10.2)) on REG from formal ontologies can be understood as generating attributive descriptions. [Section 2.6]
- Proper names pose difficult challenges to the theory of reference, because of their behaviour in modal contexts and counterfactuals contexts [Kripke, 1980] (section 2.7). Computational work on REG has usually bypassed proper names, but we shall argue in chapter 7 that they should be treated as first-class citizens. [Section 2.7]

3 The Psychology of Reference Production

Once upon a time, the dividing line between psycholinguistics and computational linguistics was sharp: psycholinguists designed and conducted experiments with human speakers and hearers, whereas computational linguists constructed algorithms and computer programs. These days, the line is not always so easy to draw, because many computational linguists – and theoretical linguists as well – now conduct experiments with human participants as well, to inform and validate their algorithms. Conversely, some psycholinguists express their models in terms of computational algorithms. Still, there are differences in outlook between these tribes, for instance because psycholinguists aim to understand how and why people communicate, whereas many computational linguists have the construction of practically useful systems as their foremost aim (see e.g., section 1.2).

This chapter offers an overview of some of the main insights in reference production that have emerged from experiments with human speakers. Far from covering this entire research area, I will focus on concepts that will be important in later chapters. These concepts include: common ground (section 3.1), audience design (section 3.2), the role of rationality and the Gricean Maxims in communication (section 3.3), and the idea that some properties appear to be psychologically more "preferred" than others (section 3.4). I shall also discuss a recent experiment that compares some of these concepts, pitting preference against one of the Gricean Maxims (section 3.5). Concluding the chapter, I will briefly highlight the role of collaboration in reference (section 3.6) and discuss the important question of ecological validity (section 3.7).[1]

3.1 Common Ground

We have seen in section 1.7 that speakers keep track of the information that they share with their audience. This theme, often treated as peripheral by computational linguists, has been explored in great depth by logicians and game theorists (e.g., [Fagin et al., 1995]) and by psychologists, and we shall briefly summarize how it has been treated by the latter group of researchers who have tended to call it "common ground", or "common knowledge". Note, for a start, that common knowledge is not simply knowledge that the speaker and hearer

1 The discussion in this chapter draws on relevant parts of [van Deemter et al., 2012b], [Paraboni et al., 2007], [Paraboni and van Deemter, 2014], [Mitchell et al., 2013c], and [van Gompel et al., 2014].

both possess. The matter can be explained by contrasting common knowledge with a weaker concept, which has sometimes been called mutual knowledge.[2]

> "Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly, 'Peter told me he will be late again', then the mutually known fact is now commonly known." [Vanderschraaf and Sillari, 2009]

In these authors' terminology, *mutual* knowledge is simply knowledge that each member of a group of two or more people possesses. *Common* knowledge (also known as common ground) might be informally characterized as knowledge that is *publicly* shared by a group of people. A and B have mutual knowledge of a proposition p if and only if A knows p and B knows p. They have *common* knowledge of p if they have mutual knowledge of p, and A knows that B knows p, and B knows that A knows p, and A knows that B knows that A knows p, and so on, *ad infinitum*. Logicians and game theorists have proposed various mathematically precise definitions of common knowledge (including cases where there are more than two knowers), which get rid of the imprecise words *ad infinitum* in different ways [Vanderschraaf and Sillari, 2009], the details of which do not matter here.

Common knowledge, in its strict sense, involves an infinite number of levels of epistemic embedding (x_1 knows that x_2 knows that x_3 knows that, *etc.*). To see why, it is worth following the reasoning in [Clark and Marshall, 1981], a piece of work remarkable not only for its cogency but also because – uncharacteristically for a psycholinguistics paper – it does not rest on experimentation but on commonsense reasoning. Let us dive into the middle of their paper, focussing on a moderately high level of epistemic embedding. They observe that, in order for an NP to be felicitously used as an RE referring to r, at least the following condition needs to hold:

> *Condition:* The hearer knows that the speaker knows that the hearer knows that NP $= r$.

It will be useful for us to agree on some terminology, so let p be a proposition that does not involve anyone's knowledge, such as the proposition NP $= r$,

2 Terminology differs across authors: what VanderSchraaf and Sillari call common knowledge, Clark and Marshall call mutual knowledge (e.g., [Clark and Marshall, 1981]).

for example. Then let us call an expression of the form "so-and-so knows that p" an epistemic expression of level 1, an expression of the form "so-and-so knows that so-and-so knows p" an epistemic expression of level 2, and so on, making the condition above an epistemic expression of level 3. Now here is the example used by Clark and Marshall to show that the Condition above is necessary. A confusion between two Marx Brothers' movies lies at the heart of the example.

> *On Wednesday morning Ann and Bob read the early edition of the newspaper and discuss the fact that it says that* A Day at the Races *is playing that night at the Roxy. Later, Ann sees the late edition, notes that the movie has been corrected to* Monkey Business*, and marks it with her blue pencil. Still later, as Ann watches without Bob knowing it, he picks up the late edition and sees Ann's pencil mark. That afternoon, Ann sees Bob and asks, "Have you ever seen the movie showing at the Roxy tonight?"*
> ([Clark and Marshall, 1981], version 4 of the basic scenario.)

Bob knows that "the movie showing at the Roxy" = *Monkey Business*, because he saw the late edition of the newspaper. Ann knows he knows this because she watched him reading. Bob, however, does not know that she knows that he knows, because he doesn't know that she watched him. How will Bob interpret Ann's utterance, "Have you ever seen the movie showing at the Roxy tonight?" If you think the situation through, you will realize that Bob – if he is rational and alert – believes that Ann assumes Bob to think of *A Day at the Races* as the referent of her NP ("the movie showing at the Roxy tonight"), not of the other movie; a misunderstanding results. In real life, Ann might add, "I watched you reading the late edition", thereby restoring the truth of the Condition.

This example involves only 3 levels of epistemic embedding, but common knowledge can break down at any level of embedding, which is why a more general Condition is needed that covers them all (as when the clause *ad infinitum* is used). Although it is possible to use increasingly complex versions of the scenario above (as Clark and Marshall did), it will be convenient to switch to a different scenario, which generalizes more easily: Ann and Bob are using a flaky electronic mail connection to arrange a meeting. Initially they agree to meet in cafe c, but then Ann realizes that cafe c' would be even nicer, and she communicates this change of plan to Bob in an email. Unfortunately only about 10% of messages arrive, so after an important email, the recipient would be wise to send an acknowledgment. This acknowledgment, however, could

also fail to arrive.[3] We focus on the RE "the cafe where Ann wants to meet" (abbreviated t), which might play a role in a discussion between Ann and Bob. The proposition whose epistemic status we shall explore is: $t = c'$.

Suppose in each situation Ann and Bob can send only one email, and all messages sent so far have actually arrived. We can sum up the first few levels of knowledge, writing "$S : yes$" for "the Speaker knows for sure that $t = c'$" and "$H : no$" for "it is not the case that the Hearer knows for sure that $t = c'$"; in the same way, HS abbreviates "the Hearer knows that the Speaker knows that", and so on. We number the messages so 1 is the email in which Ann tries to tell Bob about the change of plan and 2 is Bob's first acknowledgment.

1. H has received message 1:
S:yes. H:yes. HS:yes. SH:no.

2. S has received acknowledgment concerning arrival of message 1:
S:yes. H:yes. HS:yes. SH:yes. SHS:yes. HSH:no.

3. H has received acknowledgment concerning arrival of message 2:
S:yes. H:yes. HS:yes. SH:yes. SHS:yes. HSH:yes. SHSH:no.

4. S has received acknowledgment concerning arrival of message 3:
S:yes. H:yes. HS:yes. SH:yes. SHS:yes. HSH:yes. SHSH:yes. HSHSH:no.
Etcetera

H and S share more and more information after each message, yet there is always at least one epistemic embedding missing. At level 1, for example, we have SH:no because, before an acknowledgment has reached her, S has no way of knowing for sure that H knows that $t = c'$; for all she knows (at a 90% likelihood), H might believe that Ann wants to meet in cafe c (i.e., $t = c$). Something analogous is true at each of the four levels: the acknowledgment received by an agent a at a given level gives certainty that the last message sent by a has reached the other agent, b, but it creates a new uncertainty because b is left uncertain as to whether b's acknowledgment has reached a – unless and until this uncertainty is removed by a new acknowledgment received at the next level. This pattern can be expanded to any finite number of levels, where each new situation shows a breakdown in common knowledge, although always at an increased level of epistemic embedding (using an increment of 1).

3 My email scenario is a variant on the theme of two generals, each of whom will attack if and only if he knows that the other one will attack (e.g., [Lewis, 1969] on coordination games).

In situation 1, S believes there to be a 90% likelihood that her message has failed to arrive. Consequently, she believes there to be a 90% likelihood that Bob believes "the cafe where we have agreed to meet" $= c$ (as originally agreed) and a 10% likelihood that Bob believes "the cafe where we have agreed to meet" $= c'$ (as Ann ultimately intended). In other words, Ann and Bob may well find themselves in different cafes. The more complex situations (2,3,4, etc.) are analogous. I don't know how Ann and Bob should act in each of these, because this depends, for example, on Ann's degree of preference for c' over c. We do observe that anything short of common knowledge risks misunderstandings. Common knowledge is never attained, unless Ann and Bob pick up the phone and establish direct contact.

These arguments show that common knowledge is a highly complex kind of meta-knowledge. Clark and Marshall understood that it is unlikely that all the different levels of epistemic embedding are considered by speakers and hearers. Indeed, given that there are infinitely many such levels, of ever greater complexity, common knowledge in its strict sense is often unattainable.[4] The authors concluded that some psychological shortcuts have to exist. They briefly discuss the possibility that speakers and hearers might simplify matters by making estimations based on a limited number of levels of epistemic embedding – say, the first two or three. Their article, however, is a plea for a different set of heuristics, based on general characteristics of the situation in which REs are uttered. They observed that, in many situations, common knowledge is enforced by "triple co-presence": situations in which the speaker, the hearer, and the entities under discussion are all physically present [Clark and Marshall, 1981]. For example, consider the innocent start of Clark and Marshall's original scenario, where Ann and Bob are jointly reading the early edition of the newspaper. Had this been the end of the story, then this joint reading would have guaranteed that Ann and Bob have it as common knowledge that "the movie showing at the Roxy tonight" denotes *A Day at the Races*: all, infinitely many, levels of embedding would have been underpinned by this one single event. The beauty of this shortcut is that it does not just offer a rough approximation, but the real thing: common knowledge itself.

Clark and Marshall contrast this situation (which they call *personal* common ground with *linguistic* common ground, where information has been shared through verbal communication, and with *communal* common ground, which

4 See [Fagin et al., 1995], chapters 6 and 11, for logical and game-theoretical analysis and for applications in computing.

arises from being in a shared community of some sort (e.g., psycholinguists, stamp collectors, or people living in Aberdeen).

An example of linguistic common ground arises in the example above from Vanderschraaf and Sillari, when one of the students says, for everyone to hear, "Peter told me he will be late again", making Peter's lateness commonly known [Vanderschraaf and Sillari, 2009]. An example of communal common ground arises in Aberdeen, whose residents have direct acquaintance with a large grey building known as "Marischal College". The assumption that others in Aberdeen are similar to themselves in this respect allows them to assume that the most obvious features of the building, including its name, are in (communal) common ground. This is what allows them to point out the location of a convenience store to each other by saying "the shop right opposite Marischal College", assuming this to be understood. When speaking to someone unfamiliar to Aberdeen, they might decide to include more information to ensure that they are understood. Later research has confirmed that speakers are broadly able to distinguish between knowledge that is available to members of their own community only and knowledge that is also available to others [Jucks et al., 2008, Nickerson et al., 1987], even though there does exist a tendency to overestimate the likelihood that information known to them is known to others [Fussell and Krauss, 1992]. In chapter 13, we shall show how large knowledge resources such as the world-wide web can help us estimate the likelihood that a given fact may be in communal common ground.

3.2 Audience Design and the Egocentricity Debate

When children start communicating, one of the first things they learn is how to refer to objects [Bruner, 1983], [Matthews et al., 2007], often initially by pointing. (You point at a toy, and your mother will understand that you want it.) However, toddlers are not very good yet at understanding what exactly other people know or understand: their "Theory of Mind" – their ability to reason about other minds and common knowledge – is not yet fully developed. Well-known experiments involve a child, a mother, and a cup that is hidden inside a cupboard while the child and the mother look on. Then the mother goes away. In her absence, the cup is moved to another cupboard. Now the experimenter asks the child "When Mommy comes back, where will she look for the cup?" Below a certain age, children believe that their mother will look inside the second cupboard, where the cup actually is, but where the mother has no way

of expecting it to be. The lack of Theory of Mind evinced by these experiments shows up in children's REs, as when the child asks its mother, without further clarification, to "give me the cup".

The story of the Battle of Balaclava, with which this book opened, suggests that children are not the only ones to suffer from a lack of Theory of Mind: Lord Raglan was insufficiently aware of the knowledge of Lorn Nolan, and this caused Nolan to misunderstand Raglan's words, causing a military disaster. Some psycholinguists have argued that flaws of this kind are common, causing what is sometimes known as the *egocentricity* debate. On one side of the debate are those, including Herb Clark and Susan Brennan, who emphasize common ground and its role in human communication. On the other side is a group of researchers around Boaz Keysar who have sowed doubt about adults' ability to reason about other minds [Horton and Keysar, 1996], [Keysar et al., 2003], [Lane et al., 2006]. Let me sketch a few of their experiments before discussing one of them in more detail.

Keysar, Barr, Balin, and Brauner describe an experiment that focusses on speakers [Keysar et al., 2000]. In this experiment, which makes use of eye-tracking and other metrics, pairs of people are looking at rows of objects of different sizes. The hearer sees three candles, for example, but the speaker sees only the larger ones, because the smallest candle is hidden from his view. It is made abundantly clear to the hearer that the speaker cannot see the smallest candle. Yet, when the speaker says "get the small candle", the hearer will tend to look at the smallest *of the three* first even though, rationally, he knows that this one cannot be intended. In about a quarter of cases, the hearer's hand starts moving towards this impossible referent, and in about three quarters of these cases, the hearer actually moves this object instead of the "correct" one; in the remaining cases (about 5% in total), the initial hand movement gets "corrected", and the hearer reaches for the correct referent. In other words, hearers sometimes behave "egocentrically", as if they had no Theory of Mind. Keysar and colleagues have argued that Theory of Mind may be a bit like a fancy espresso machine that you have been given as a present: you own the machine, yet you may not use it very often [Keysar et al., 2003]; in the same way, the authors argue, adults possess the ability to reason about other people's knowledge, yet we often fail to use this ability.

A number of experiments in this "egocentric" tradition have focussed on speakers rather than hearers, making these experiments particularly relevant for the present book. An experiment, reported in [Horton and Keysar, 1996], which puts speakers under time pressure, compares speakers in two different

conditions. In both conditions, a speaker and a hearer are both looking at two computer screens. What they see is, for example, a circle that moves slowly from one screen to the next. The speaker is asked to describe what she sees. In both conditions, she sees one other circle on her screen, of a different size than the target, which allows her to think of the target as "the small circle" or "the large circle", depending on the size of the other circle. In one condition, the hearer sees exactly the same as the speaker. In the other condition, the hearer sees *only* the target circle, without another one to compare it to; this makes expressions like "the large circle" incomprehensible to the hearer, because the degree adjective "large" is meaningless to him. It is made abundantly clear to the speaker that the hearer can see only one circle. Therefore, speakers with good Theory of Mind would be much more cautious in the informationally asymmetric condition than in the condition where speaker and hearer see the same things. Yet, the authors found that there was essentially no difference between the (very substantial) number of degree adjectives used in the two conditions. Expressions such as "the large circle" were used equally often in both situations. Speakers, in other words, did not make any allowance for the fact that hearers were unable to make size comparisons.

The amusingly titled article *Don't Talk About Pink Elephants!: Speakers' Control Over Leaking Private Information During Language Production*, by Lane, Groisman, and V. Ferreira, reports on an experiment that focusses on the same phenomenon, asking to what extent speakers are able to control their Theory-of-Mind use [Lane et al., 2006]. Speakers were shown four shapes, one of which was occluded from the view of the hearer. In the crucial condition, the shapes presented to the speaker included a smaller or larger version of the target shape. Two conditions were used: a baseline condition in which participants were simply asked to refer, and a test condition in which they were asked to refer *without providing addressees with information about the hidden shape*. Speakers in this condition, in other words, were instructed not to "leak" private information. Curiously, the authors found that information was leaked *more* often in the test condition! It appears that when speakers' attention was drawn to the forbidden shape, they became more likely to mention it, and this is what led the authors to the title of their paper.

Another study has given rise to an interesting debate. The study focussed on speakers' choices between names and descriptions [Wu and Keysar, 2007]. Participants were shown unfamiliar complex shapes and taught equally unfamiliar names for these shapes (e.g., one shape was called *Abypit*). Shapes were

named consistently, so speakers and hearers would never learn to associate different names with the same shape. However, some objects were not named, so these objects could only be described, not named. Crucially, not every participant learned the same number of names, so some shapes could be named by some but not all participants. Participants were grouped in pairs, each of which contained a speaker and a hearer. The speaker's task was to allow the hearer to single out the target shape (which the speaker saw on his monitor) on a monitor that showed three shapes, namely, the target shape and two distractor shapes.

Pairs of speakers/hearers were distributed over two conditions. In one condition, pairs shared a name for as many as 60% (high overlap) of shapes; in the other condition, this was only 20% (low overlap). The experimenters were curious to see whether this information about the likelihood of sharing a name would affect speakers' choice between names (e.g., "Abypit") and descriptions (e.g., "circle on the top, and then two sort of arrows, so that makes it a little neck"). Names, if shared, are highly efficient, so when names are almost always shared, one might expect them to occur frequently, even in those situations where speakers could have realized that the name was unknown to the hearer. And this was indeed what the authors found: high-overlap speakers used three times as many privileged names (i.e., names that the hearer had not learned) as low-overlap speakers.

Wu and Keysar appeared to give new ammunition to researchers who stress the "egocentric" perspective, because once again (cf., [Keysar et al., 2000] and [Keysar et al., 2003]), a situation had been identified in which common ground takes a backseat. A hint of paradox adds interest to these findings because sharing *more* information (i.e., being in the high-overlap condition) also meant experiencing *more* misunderstandings. Conditions of high overlap appear to invite risky behaviour.

However, in a follow-up to [Wu and Keysar, 2007], Heller and colleagues took a close look at Wu and Keysar's experiment, making it the basis for a new study whose aim is to find out whether the cases in which Wu and Keysar's speakers used names for privileged shapes were actually egocentric [Heller et al., 2012]. First of all, the authors wondered whether the use of names in these situations was as risky as the authors of the study interpreted them to be. To find out, they divided these utterances into five categories:

Name alone (e.g., "Uhm cortlog")

Name-then-description (e.g., "inta, you havent seen it, its four arrows")

Description-then-name (e.g., "a box and triangle, its a molget is the name of it")

Description-with-name (e.g., "it looks like the chicapee one except with a long tail")
Description (e.g., "Um its two triangles kissing")

Assigning utterances to these five categories, it turned out, first, that in situations where the hearer knew the name, most of the utterances with names were *Name alone* utterances. But, second, in situations where the hearer did not know the name, *Name alone* utterances were rarely used: separating between High Overlap and Low Overlap, the figures were: Shared (High 0.64 vs. Low 0.48), Privileged (0.05 vs. 0.01), and New (0.05 vs. 0). Shared situations are defined as ones in which the name is known to both speaker and hearer, Privileged ones are ones where only the speaker knows the name, and New situations are ones where neither knows the name.

The *Name alone* form was used more often in Shared situations than in Privileged ones, but Privileged and New situations did not differ. If one focusses on uses of names that do not contain additional descriptive content (which could potentially help the matcher identify the intended referent), then these were basically limited to situations where they would be understood. Given these new findings, it is suddenly starting to look as if Wu and Keysar were wrong to focus on names without looking at the way in which they were used.

Having found that Wu and Keysar's speakers appeared to be well aware of the distinction between shared and privileged information, Heller and colleagues wondered whether hearers might be able to understand by listening to these speakers' utterances, which ones were uttered by someone who believes the name to be known by the hearer. So, the authors asked naive hearers to judge, for every name in the speakers' spoken utterances, whether they thought the speaker assumed that the hearer knew that name. The results were striking: of the *Name alone* trials, 97% were judged as assumed shared and only 1% were judged as assumed privileged. Of the *Name-then-description* trials, by contrast, only 14% were judged as assumed shared and 86% were judged as assumed privileged. Why names were used as often as they were in cases where the hearer wouldn't have come across them before is not entirely clear, but the authors sensibly hypothesize that this was done to "teach" the hearer the name, in case the shape in question would come up in the future.

The jury is still out on many of the questions in this area. Even before Heller et al.'s results, egocentric findings had been questioned. Some argued, for example, that egocentricity may be an artefact of a particular type of experimental setup [Brown-Schmidt, 2009]. Heller and colleagues have added to these questions, yet the work of Keysar and colleagues makes it plausible

that Theory-of-Mind use is subject to important limitations, particularly when participants are under time pressure [Horton and Keysar, 1996]. Much is still unknown about the depth and sophistication of speakers' and hearers' modelling of each others' knowledge and how this relates to established findings about "thinking fast" (cf., [Kahneman, 2012]).

3.3 Rationality and the Gricean Maxims

In section 2.8 we briefly introduced the Gricean Maxims of Quality, Quantity, Relation, and Manner. Even though these Maxims have attracted considerable interest from psychologists, and despite some promising studies, it is not yet clear to what extent Grice's claims in this area hold up to psycholinguistic scrutiny [Bott and Noveck, 2004], [Breheny et al., 2006], [Huang and Snedeker, 2009]. In a survey article, July Sedivy wrote the following about the relationship between real-time language processing and conversational implicature: "(...) systematic study of this relationship is still in its very early stages and (...) we are very far away from having a good general understanding of the nature of pragmatic processing." [Sedivy, 2007]. Despite a recent upsurge in psycholinguistic work on issues in linguistic pragmatics, Sedivy's assessment appears to be valid still. In particular, psycholinguists are unclear which types of implicatures are "calculable" in actual language use. We leave these issues aside here, turning to the question of what the Gricean Maxims say or imply about reference. Precisely this question was the starting point of [Dale and Reiter, 1995], which has long dominated computational REG. The consequences of this article will occupy us for much of the next two chapters, but first we shall offer a general perspective on what the Maxims might mean in connection with reference. We discuss them one by one.

The Maxim of Quality. Few computational or psycholinguistic studies have focussed on the Maxim of Quality. The philosophical literature, however, abounds in example utterances in which Quality is, or appears to be, breached. Some of these were discussed in the previous chapter. In one type of breach, the speaker ascribes a property to the referent that is not true of the referent but of something associated with it. This happens when a waitress says (in an example due to Nunberg) "The ham sandwich is getting restless", referring not to a consumer product but a person [Nunberg, 1978]. Only the most pedantic would call such utterances incorrect: their apparent incorrectness can be

explained away as using metonymy in order to shorten an otherwise cumbersome description ("the person who ordered a ham sandwich").

In another breach of Quality, someone might say *The man with the Martini is happy*, where the intended referent drinks wine instead of Martini (section 2.6). Different explanations are possible: perhaps the speaker was honestly mistaken, in which case Quality was not breached after all; or the speaker was unsure whether the drink was wine or Martini, and realized that as long as the look of the drink is consistent with it being a Martini, the description *the man with the Martini* will work well. The speaker might reason (using remarkably complicated Theory of Mind; cf., section 3.2) that the hearer is likely to be as ignorant of the content of people's glasses as she is and that, confronted with the utterance "the man with the Martini", the hearer must scan the room for people who *may* be thought by the speaker to be drinking a Martini. The conclusion that these utterances breach Quality, however, is difficult to avoid.

The Maxim of Quantity. Quantity has always played a central role in psychologists' thinking about reference (e.g., [Olson, 1970]). Its application to reference seems straightforward: REs should contain the minimum amount of information that makes the identification of the referent possible. Before we proceed, how should "amount of information" be measured? Should properties simply be counted, or should their (implicit) internal structure be taken into account? Consider being a bachelor. Suppose there is only one unmarried male in the domain, who happens to be adult. Does this make the RE "the unmarried male" nonminimal, because the expression "the bachelor" would have sufficed? Or should one argue in the opposite direction, that BACHELOR contains *more* information (given that a bachelor must be adult), concluding that it is "the bachelor" that is overspecified? The literature contains little discussion of such matters, despite the centrality of the notion of brevity, so the question must be left unresolved here (cf., section 5.9).

```
UNMARRIED,MALE   >   BACHELOR
      =                  =
UNMARRIED,MALE   <   UNMARRIED,MALE,ADULT
```

Grice's Maxims differ in interesting ways. Quality, for instance, avoids requiring that speakers speak the truth: they should merely avoid saying "what you believe to be false". The Maxim of Quantity, by contrast, does not contain such a hedge: it requires that speakers should not give too much information; a more cautious formulation would ask merely that speakers avoid producing what they *believe* to be too much information. Perhaps, given the limitations of human perception and reasoning, this cautious formulation would have been preferable. Years of research on human information processing, including the Nobel prize-winning work of Kahneman and Tversky, has taught us how, in a

wide variety of situations, people rely on "cheap heuristics" rather than careful thought (e.g., [Kahneman, 2012]). In view of this large body of work, it would be a miracle if people's spontaneous speaking allowed them to always flawlessly calculate the minimum number of properties that identifies a referent. For although a referent can stand out from a large set of distractors, if the referent differs from all distractors in terms of one and the same property – for example, the referent might be the only red shape in a sea of green ones – this "pop-out" effect is far weaker if the referent differs from some distractors by one property (e.g., its colour) and from others by another property (e.g., its shape) [Treisman and Gelade, 1980].

Experiments with human speakers (starting with [Ford and Olson, 1975], [Whitehurst, 1976], [Pechmann, 1989]) consistently show up considerable percentages of overspecified REs – defined as REs from which one or more properties can be removed without causing ambiguity regarding the intended referent of the RE – and a few percentage points of underspecified REs, which fail to identify the intended referent. Thus, even if "full" brevity were something speakers *try* to achieve, they do not always succeed. Where they produce a non-minimal RE, it would be misguided to draw Gricean implicatures (i.e., inferring that they intend to convey additional information), because such implicatures rest on the assumption that deviations from the Maxims are intentional.

Many researchers have taken these insights onboard. A useful concept in this connection is the *Discriminatory Power* of the properties considered for inclusion in an RE. Suppose M is a set of domain elements not yet ruled out, P is a property defined over M, and r is the referent. We abbreviate the set of distractors, $M - \{r\}$, as *Dis*. Now we take the Discriminatory Power (DP) of P to be the number of distractors removed by P as a proportion of the total number of distractors (cf., [Dale, 1992], section 5.2). This can be written as follows, where $\|.\|$ stands for the number of elements in a set. As before, $[\![.]\!]$ stands for the extension of a property.

$$\text{DP}(P, M) = \frac{\|[\![\overline{P}]\!] \cap Dis\|}{\|Dis\|}$$

Consider the example domain of Table 3.1 (repeated from chapter 1). Let "zoo" be the property of being an animal in the infirmary of this zoo. Now if the referent is a lion, we have DP(lion, zoo) = $1/2$ (because the property "lion" removes 1 of the 2 distractors). If the referent is a tiger, then DP(tiger, zoo) = 1 (because "tiger" removes each of the 2 distractors). The idea can also be used when the context has already been narrowed down. For example, in the context C that results from using the property "lion", DP(Kenya,C) = 1. The exact role

IDENTIFIER	SPECIES	ORIGIN	WEIGHT	INJURIES
a	*lion*	*Kenya*	$102kg$	*paws, teeth*
b	*lion*	*India*	$100kg$	*paws*
c	*tiger*	*China*	$310kg$	*back*

Table 3.1

Information about the animals in a zoo, shared by speaker and hearer.

of Discriminatory Power can vary, but one possibility is to say that where there is a choice between properties that do not have equal DP, the property with the highest DP should be chosen. We shall see in chapter 4 how this idea can be interpreted computationally.

Recent experiments have investigated the impact of overspecification on hearers' and readers' comprehension of REs. [Engelhardt et al., 2006], for example, presented experiments in which speakers overspecify almost one-third of the time; interestingly, hearers do not judge overspecifications to be any worse than minimal descriptions. On the other hand, the paper also reports on an eye tracking study showing that overspecification causes comprehension to take longer in some types of cases. Similar findings were reported in [Engelhardt et al., 2011], where overspecified descriptions (as in "look at the red star" in a context in which there is only one star) took longer to be interpreted than minimal descriptions("look at the star").

An overspecified RE can be useful because it contains a (logically superfluous) property that allows the hearer to focus on a particular part of his visual field. [Arts, 2004], for example, who let speakers refer to buttons on electronic equipment, found that when a noun alone would suffice for disambiguation (as in "the button"), *lower* recognition times result if speakers use spatial overspecification (as in "the button *on the top*"). Similar things were found in [Paraboni et al., 2007] and [Paraboni and van Deemter, 2014]. By showing that speakers prefer the same overspecified expressions as hearers [Paraboni et al., 2007], this work suggests that, in the situations at hand, human speakers take the hearer's perspective seriously, designing their descriptions in such a way that certain types of problems for their audience are avoided.

The effects of overspecification on the recipient of an RE are not yet well understood. Of particular interest, at the moment, is the question of whether speakers tend to produce the kinds of overspecified REs that have benefits for hearers (and whether they do it for the benefit of their hearers: this is known

as the question of Audience Design). After all, if overspecification is merely an unintended consequence of speakers' cognitive limitations, then there is no reason to assume that the resulting expressions are of particular benefit to the recipients of their REs. We shall return to these difficult issues in chapter 12, where some of the experiments mentioned above are described in more detail.

Communication involves risks [Carletta and Mellish, 1996a]. The primary risk associated with REs is for the recipient to misunderstand what the intended referent of an RE is. A key finding of Kahneman and Tversky's *Prospect Theory* is that people tend to be risk averse: experiments reveal that most of us prefer the certainty of a small loss over a small chance of a more substantial loss, even if the probabilistically *expected* loss is greater in the former case than in the latter (in which case standard decision theory would predict the latter choice). These results are well established across a large area of tasks. Suppose now that verbosity (i.e., the use of a non-minimal description) implies a small loss; this seems reasonable, given that verbosity takes time and effort. Suppose, furthermore, that referential misunderstandings imply a substantial loss. Then Prospect Theory predicts that speakers will tend to prefer overspecification (certainty of a small loss; possibility of a large gain) over underspecification (certainty of a small gain; possibility of a big loss). Broady speaking, this prediction is borne out by the facts.

The Maxim of Relation. Relevance has rarely been studied in connection with reference, except by theoreticians (see section 2.2), who were aware that the properties that constitute an RE are not chosen for their denotation alone. A study on weather predictions went into some computational detail (cf., chapter 14), where this work will be discussed): the issue came up when describing where on the map a particular weather phenomenon occurred: "Changes in precipitation type are more commonly seen in higher elevation areas where the air temperature is generally lower, so a spatial description of such an event should make use of a reference frame that reflects this interaction" [Turner et al., 2008]. Hence, to say "Precipitation was high in urban areas" would be odd, and it is better to describe this area in terms of a different reference frame, such as elevation. Turner et al.'s position is that, where possible, properties should be mentioned that have a causal bearing on the proposition expressed. For them, in other words, relevance meant causal relevance. Section 16.2 will discuss other elements of a computational solution to the problem of relevance in REG.

The Maxim of Manner. This Maxim advises against anything that can make an expression cumbersome or unclear. The Maxim warns against verbosity, against syntactic and lexical ambiguity, and against the kind of messiness that can make a text difficult to understand. Focussing on reference, the most obvious implication is that clear words and unambiguous syntactic constructs should be chosen to express the logical content of an RE. Issues of this kind have been studied far less often in REG than issues of Content Determination (cf., chapter 1), but some recent work has started to explore these issues, asking how ambiguity should be understood in this context and how clarity (advocated by the Maxim of Manner) should be traded off against brevity (advocated by the Maxim of Quantity) ([Khan et al., 2012], see section 8.8).

The Gricean Maxims will play a role throughout this book, and at the end of the book, we shall revisit them to see how they should be viewed in light of our present understanding of reference (section 16.2). But despite their prominance, the Maxims alone do not suffice to understand reference: rationality alone is not enough.

3.4 Intrinsic Preference for Certain Attributes

In some areas of psycholinguistics, computational models have become a well-established tool. An example is *lexical access*, which seeks to predict what word will be triggered in a speaker by a picture of an object and how quickly (see e.g., [Dell et al., 1997], and the Weaver++ model of [Levelt et al., 1999] for computational models). However, until recently psycholinguists seldom constructed computational models of the process on which we focus in this book. For example, the results discussed in section 3.2 were never expressed by means of a computational model. By and large, they have left the construction of algorithms to computational linguists and language engineers.

But imagine you had to outline, in broad algorithmic terms, how the semantic content of REs can be determined, aiming to find a semantic content in each situation that resembles the kind of content that a human-produced RE might possess. Suppose that all you had to go by was the Gricean Maxims. How would you do it?

Painting with a broad brush, the Maxim of Quality would tell us to find all those properties that can truthfully be ascribed to the referent; call this set of properties S. Relation would advise us to select from S the properties that are relevant to the situation in which the utterance is made, resulting in some

$S' \subseteq S$. Quantity would tell us to choose from S' a minimally sized subset $S'' \subseteq S'$ that manages to single out the referent. Manner, finally, would tell us to be careful when conveying this combination of properties S'' in actual words, avoiding ambiguity, verbosity, and other problems that can arise at this level. Could this possibly be the right account?

A long tradition of psycholinguistic studies says that this cannot be the whole story. In a nutshell: the Maxims are not enough, because there are things that the Cooperativity Principle (which the Maxims are designed to implement) does not capture. These additional things have little to do with what is rational or cooperative *a priori*, but everything with the peculiarities of the human mind. Let me summarize some of the findings in this area.

Pechmann, when focussing on visual domains, made use of the well-established idea that properties can be clustered into closely related groups that share the same "attribute". For example, each of the properties red, green, and brown can be seen as values of the attribute COLOUR, predicting that they play similar roles in communication and that they might be processed in similar ways by the human brain. Attributes can be thought of as implementing the philosophers' notion of a conceptual perspective, which we discussed in section 2.2. (Diagrams such as 3.1 assume attributes as well when they use labels such as the SPECIES and ORIGIN of an animal in the zoo.)

Pechmann claimed that perceptually *salient* attributes, such as colour, tend to be selected for inclusion in an RE *before* other attributes, such as size, causing them to be used even when they have no contrastive value. They appear to be intrinsically (i.e., because of what they are, and not just in terms of what they achieve in a particular situation) "preferred", an established word that we shall continue to use despite its unintended overtones of intentionality. Later work confirmed that speakers, faced with a visual domain, have a strong tendency to use the COLOUR of the referent (Schriefers and Pechmann, 1988; Pechmann, 1989, Viethen et al., 2012), frequently resulting in overspecification and, consequently, an infringement of the Maxim of Quantity. Size is common as well, particularly when the domain contains several salient objects of the same colour [Brown-Schmidt and Tanenhaus, 2006].

At least in English, one attribute stands out as even more highly preferred than colour, namely, the TYPE of the referent. The notion of a type is not easy to define, but it captures the intuition – going back at least to Aristotle's *categories* – that although a referent may have various properties (black-and-white, large, possessing a short tail and sharp teeth), at a more general level it is a particular *type* of thing. Linguistically, types are often encoded as nouns, and

this fact in itself might add to their centrality. Consider Fido, a large black-and-white Scottish sheepdog. REs typically contain a TYPE, even when none is required for singling out the referent: we may call Fido "the black outdoor *dog*", for example, even in situations where the Logical Form {black, outdoor} would suffice to identify the referent. Once again, an Intrinsic Preference for an attribute (i.e., the TYPE attribute in this case) causes infringements of the Maxim of Quantity.

If something is a sheepdog, it must be a dog as well, of course, and a mammal and an animal. There is often one particular level along this hierarchy, known as the *basic level*, which is psychologically most important (e.g., [Rosch, 1978]) and contributes most to the *Gestalt* (i.e., the mental representation) of the referent; in the example at hand, DOG is probably at this basic level. Basic-level nouns are learned by children first, and they are manipulated with greater speed than other nouns, for example as reflected by subjects' response times following questions like "Is this a so-and-so?" Some nonbasic values may be so dispreferred that using them would make a strange impression. For example, it would be odd to call Fido "the black outdoor *entity*".

Why some attributes are more highly "preferred" than others is an interesting question. The centrality of *types* is understandable on functional grounds, because they are crucial in affording inferences about the referent. In animals, types are often species. Cats and dogs, for instance, may differ along a number of dimensions, but the fact that some animals are cats and others dogs (from which other differences follow) seems more fundamental than all others. In other cases, the root cause of the preference degree of an attribute may be different. It has been suggested, for example, that size is less preferred than colour because the assessment of size requires more cognitive effort, given that it requires comparison to other objects, something that is less obvious for colour. Colour appears to be central to the speaker's mental representation of an object in a way that most other properties are not. Colour is thought to be the first property that our visual system processes, followed by size [Murray et al., 2006], [Fang et al., 2008], [Schwarzkopf et al., 2010]. These findings have been interpreted in terms of the *codability* of an attribute, that is, the ease with which that attribute can be included in a mental representation of an object [Belke and Meyer, 2002].

Despite the evidence behind Intrinsic Preference, there are reasons for caution. Decades ago, [Hermann and Deutsch, 1976] showed that the size of the contrast between the referent and its distractors matters. In experiments involving candles of different heights and widths, for example, if the referent is both

the tallest and the fattest candle, subjects tended to say "the *tall* candle" when the tallest candle is much taller than all others while the same candle is only slightly wider than the others; if the reverse is the case, speakers switch to "the *fat* candle". Later, Sedivy found that when the colour of a referent is predictable – as when we speak about bananas, which are normally yellow – then speakers' inclination to use colour drops sharply [Sedivy, 2003]; conversely, attributes that are normally dis-preferred can become preferred when they are situationally meaningful. Recent experiments show that when the difference in size between the referent and its distractors is huge, speakers no longer "prefer" colour over size [van Gompel et al., 2014] (see also section 6.3). In short, Intrinsice Preference is not absolute. Soon we shall see how Intrinsic Preference has affected REG algorithms (section 4.6); at the end of the book we shall review these issues in light of all the evidence (section 16.2).

I have discussed Intrinsic Preference separately because it is different from the Gricean Maxims. Admittedly, given a hearer's preference for certain attributes over others, she may well produce and comprehend such REs with particular ease, and her hearers may share this trait. This will mean that speaking in accordance with Intrinsic Preference fits the Cooperativity Principle because hearers are attuned with it. If this is the complete story, however, there is nothing inherently cooperative in taking Intrinsic Preference into account; one Preference Order is as good as the next one, as long as all speakers share it – a bit like driving on the left side of the road.

3.5 Comparing Preference with Discrimination

We have identified a number of factors that affect reference production. The Intrinsic Preference for certain attributes over others was the topic of the previous section; the others can be derived from the Cooperativity Principle:

a. Truthfulness (related to the Gricean Maxim of Quality)
b. Discriminatory Power (related to the Gricean Maxim of Quantity)
c. Clarity of Formulation (related to the Gricean Maxim of Manner)
d. Relevance (related to the Gricean maxim of Relation)
e. Intrinsic Preference or Codability

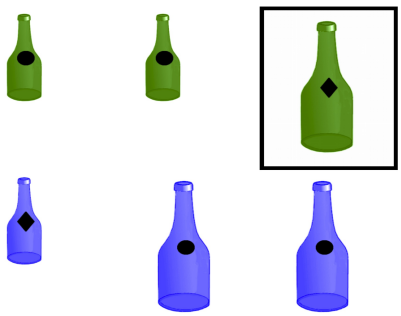


Figure 3.1

A domain as shown to participants in the experiment of [Gatt et al., 2013a]. The target referent appears in a black frame. The black shapes on the bottles are called "patterns".

We have seen that conflicts can arise among these factors. Let us focus on Intrinsic Preference and Discriminatory Power, the two factors that have influenced computational work more than any others (see chapters 4 and 5), summarizing a study reported in [Gatt et al., 2013a]. My co-authors and I were aware of the importance of Intrinsic Preference and wanted to know whether the DP of a property really does play an additional role. We performed a study in which the strength of the latter factor was varied. Subjects were shown domains containing 6 objects, which differed in terms of their colours, sizes, and/or patterns, as in Figure 3.1.

Colour was chosen because, as we have seen, it is thought to be highly preferred. Size was chosen for the same reason, and because it is assumed to be less highly preferred than colour. Additionally, we chose pattern (implemented as a black mark on the face of the bottle) because we were looking for an attribute that is easily visible yet unlikely to be highly preferred. Domains were selected in such a way that each bottle could always be minimally identified using any two of the three attributes (i.e., colour and size, colour and pattern, size and pattern). There were three conditions, depending on which of the three attributes has the highest DP. In the domain of Figure 3.1, for example, colour and size would each remove 3 of the 6 distractors, resulting in a DP of $1/2$, whereas pattern removes 4 of the 6 distractors, a DP of $2/3$.

Analysis of the data revealed that colour, size, and pattern were used with very different frequencies, but that the proportions of descriptions that contains colour, size, and pattern was almost completely unaffected by the question of which of these three had the highest DP. The experiment, reported in

	Colour	Size	Pattern
MDP = Colour	0.99	0.73	0.57
MDP = Size	0.99	0.76	0.55
MDP = Pattern	0.99	0.73	0.59

Table 3.2

Proportion of descriptions containing colour, size, and pattern (columns) in each condition (rows). MDP *stands for "most discriminatory property", that is, the property that rules out the largest number of distractors in a given situation.*

[Gatt et al., 2013a], does not tell us why colour, size and pattern behave so differently, but it does demonstrate that there are crucial differences between them that have nothing to do with what their DP is in the situation at hand. Traditional experiments of the kind discussed so far in this chapter can be instructive, because they answer a precise question. It is difficult to see, however, how they can give us a generic insight into the role of different attributes. The study discussed in this section is a case in point. It tells us something about the likelihood that an RE displays a certain feature (e.g., expressing a value of the COLOUR attribute), given a certain kind of domain. There are, however, things these studies cannot do:

1 They cannot predict how "preferred" a given attribute or property is.

2 They cannot tell us how Discriminatory Power and Intrinsic Preference should be combined or traded off against each other.

3 They cannot tell us, *for all domains and all referents* in them, what attributes to use.

Later chapters will use a computational approach in an attempt to get closer to answering these questions. This will be done by counting frequencies of attribute uses in specifically designed corpora (addressing point 1), formulating algorithms that propose specific ways to produce REs in specific situations, and organizing open competitions in which these algorithms are experimentally compared (addressing points 2 and 3).

3.6 Insights from Dialogue

Even though our emphasis is on production of "one-shot" REs, studies based on dialogue can teach us a lot. Up to about 15 years ago, most psycholinguistic work on reference was concerned with either the speaker or (more often) the hearer, because it is difficult to perform tightly controlled experiments that involve both. But as pointed out in [Brown-Schmidt and E.Konopka, 2011], the increasingly sophisticated and affordable use of new experimental techniques based on eye-tracking has changed this. Some of the resulting work has focussed on spontaneous dialogue, for instance by letting a participant interact with another person, who is a *confederate* of the experimenters; in this way, the participant can be observed in a setting whose parameters can be controlled in detail, yet she is involved in a genuine dialogue.

The study of reference in dialogue are starting to show, as we saw in section 1.6, that dialogue partners often create REs *jointly*. Moreover, participants align their utterances with each other in terms of the form and content of the utterances [Pickering and Garrod, 2004], [Brown-Schmidt and Tanenhaus, 2004], varying from the phonetic and prosodic properties of their speech to the syntactic structures and the words that they use and the way in which they organize their ideas. The effect on the content of an RE is summarized in [Arnold, 2008] (drawing on [Glucksberg et al., 1966], [Clark and Wilkes-Gibbs, 1986b], and on experiments in [Fussell and Krauss, 1989], [Brennan and Clark, 1996], and [Metzing and Brennan, 2003]):

> "In Clark and Wilkes-Gibbs's (1986) classic tangram study, pairs of participants performed a referential communication task (see also Fussell & Kraus, 1989; Glucksberg, Krauss, & Weisberg, 1966), in which one participant described geometric shapes so their partner could put them in order. As pairs of participants repeated the task, they developed shared terms of reference. This facilitated later trials, in which descriptions became increasingly shorter and more effective. Brennan and Clark (1996) proposed that speakers and addressees form conceptual pacts – an implicit agreement about how to conceptualize objects – and that these conceptual pacts determine their references to objects that can be described in multiple ways, for example pennyloafer or shoe (for comprehension evidence see Metzing & Brennan, 2003)." [Arnold, 2008]

Regarding attributes other than TYPE, it has been found that speakers' tendency to reuse an attribute that has featured in an earlier RE can be strong enough to overwhelm Attribute Preference [Goudbeek and Krahmer, 2012].

Dialogue-based studies are placing earlier findings in a broader perspective. One example is the finding, in [Beun and Cremers, 1998], that a distractor may be disregarded when the meaning of the sentence makes it an unlikely referent. For example, speakers happily said "the yellow block" when several yellow blocks were clearly in evidence, as long as the intended block was the only one that could be picked up (without first removing objects on top of it); hearers had no difficulty comprehending utterances of this kind. Another striking example is the finding, in [Sedivy et al., 1999] (also discussed in [Brown-Schmidt and E.Konopka, 2011]), that hearers are highly sensitive to what we have called Intrinsic Preference. We conclude this section by summarizing this remarkable study in more detail.

Imagine conditions A and B; in the former, the attribute of interest was a gradable adjective like "tall", which is not very highly preferred; in the latter, the attribute was a colour adjective such as "yellow", which is highly preferred. In both conditions, the adjective was used when the adjective was applicable to the referent and another object; also, in both conditions, a third object was present, to which the adjective could not be applied. Thus in condition A, with the RE "The tall pitcher", the scene could look as follows:

```
tall pitcher
tall glass
small glass
```

The experiment used a visual world paradigm, in which eye-tracking was employed to find out how participants were scanning the scene. The authors found that after hearing the word "tall", hearers paid more attention to the tall glass than to the tall pitcher, presumably because they know that an adjective like "tall" is not used superfluously (as would have been the case in "the tall pitcher"). In condition B, the RE could be "The yellow pitcher", in which case the scene could look as follows:

```
yellow pitcher
yellow glass
red glass
```

The two situations are structurally the same, with "yellow" playing the same role in B as "tall" in A. Yet hearers processed the RE differently: after hearing

"yellow", hearers did not pay more attention to the yellow glass than to the yellow pitcher. Apparently, hearers are attuned to the fact that colour terms can be used for other purposes than ruling out distractors, whereas gradable adjectives cannot, and they are able to use this to speed up their comprehension of REs, using the type of the adjective as "an early cue to the speaker's referential intent" [Brown-Schmidt and E.Konopka, 2011]. To put the icing on the cake, Sedivy and colleagues found that when they turned to a third condition C, which resembled B except that the referent had a predictable colour (e.g., it was a yellow banana), the colour adjective behaved as if it were not preferred. Thus, when hearers saw a visual scene representing the following situation:

```
yellow apple
yellow banana
brown banana
```

and heard "the yellow ...", they looked more at the yellow banana than at the yellow apple, presumably because they are attuned with the fact that "predictable" colour terms are only used when their use aids comprehension.

3.7 Ecological Validity of Experiments

Before moving on to the computational study of reference production, it is important to highlight an area of debate that has accompanied research in experimental psychological at least since the 1940s (e.g., [Brunswik, 1943]), namely, the question of "ecological" validity. Generally speaking, an experimental study of a particular type of behaviour is ecologically valid to the extent that it gives us information about that type of behaviour as it occurs in *real life*, as opposed to merely the type of situation created by the experimenters. Ecological validity is now understood to come in several distinct dimensions, including the experimental setting, the stimuli, and the task of the experiment [Schmuckler, 2001]. Applied to our subject of study:

experimental setting and task: are the situation in which participants are placed, and the task they are given, similar to the ones people are most likely to encounter in real life? – Or are they contrived and unusual?

experimental stimuli: are the referent, the distractors, and the domain of which they are a part, naturalistic? – Or are they somehow artificial and unlike the ones most commonly occurring in real-life language use?

Other things being equal, the more ecologically valid a study on reference production is, the more able it should be to shed light on reference production in real life. However, a strong emphasis on ecological validity can sometimes reduce the amount of *control* that the experimenter has over the situation, because it can mean placing participants in realistic situations that are so complex that it is not feasible to control all their facets. There is often, in other words, a trade-off between ecological validity and control over the experiment.

Traditionally, most experiments on reference production have stressed control. More specifically, they have tended to focus on artificially stylized (sometimes purely geometrical) shapes in a domain with a handful of distractors, involving a reference task in which reference is isolated from the participants' wider goals and actions: in essence, they are simply asked to refer to a given object, for no particular reason. The study of section 3.5 illustrates each of these shortcomings. In later chapters of the book, however, we shall encounter studies that aim for greater ecological validity.

3.8 Summary of the Chapter

We have discussed some of the main themes emerging from years of research on human reference production:

- An important concept is that of *common ground*, a key component of Information Sharing (cf., section 1.7). In its simplest form, this concept is used as follows: an RE of the form "the X" can only be uttered successfully if the information that there exists only one X in the common ground is shared by the speaker and hearer. [Section 3.1]
- It can be difficult to determine whether a given proposition is in common ground, for example, under time pressure [Horton and Keysar, 1996], or when communicating with an audience that is largely unknown. [Section 3.2] We shall turn to this problem in chapter 13.
- A number of notions that govern the production of referring expressions can be grouped under the Gricean Maxims. These we have called Truthfulness, Discriminatory Power, Clarity of Formulation, and Relevance. Different interpretations of these notions are possible, however, and conflicts between them are frequent. [Section 3.3]

- The notion of Intrinsic Preference is of at least equal weight. However, a given attribute is not always equally preferred (e.g., a large size contrast may be more preferred than a small colour contrast), and attributes may be preferred to different degrees depending on the referent. [Section 3.4]
- Computational models discussed in later chapters will aims to offer precise versions of these ideas.

At a general level of research methodology, this chapter has shown how different disciplines of Cognitive Science have started to merge: the notion of common ground, widely studied in both Formal Logic and Game Theory, has been approached here from the viewpoint of psychology; a computational and statistical perspective will be brought to bear in chapter 13. Similar observations can be made about the Gricean Maxims, which originated in the philosophy of language (recall section 2.8), and which have been investigated by psychologists and Computational Linguists alike (cf., section 3.3).

II SECOND PART: SOLVING THE CLASSIC REG PROBLEM

4 Getting Computers to Refer

Present-day computational research on Referring Expressions Generation (REG) differs significantly from work done in the 1970s and 1980s, both in terms of the questions that are asked and the methods that are employed in order to answer them. This chapter will start with a summary of the earliest REG programs, after which we focus on the phase, between 1985 and 1995, in which Robert Dale and Ehud Reiter embarked on a new direction, focussing on a simpler version of the REG problem.

We shall show how, once the problem had been framed in a novel way, new research questions emerged: Do human speakers approach it as essentially a logic puzzle, trying hard to find the most effective way to individuate a referent? Or do they rely on quick-and-dirty heuristics, oblivious to the disadvantages that these heuristics might have? The present chapter will draw the battle lines.[1] Efforts to decide these questions using experiments with human participants are left for the next chapter. Later algorithms, which address a wider class of reference problems, are discussed in Parts III and IV of the book.

We begin by digging up some of the oldest computational work on REG, starting with Winograd and Davey in the 1970s (section 4.1), followed by Appelt and Kronfeld in the 1980s (section 4.2). Then we turn to the work of Dale and Reiter, in the 1990s, which resulted in the classic REG task (sections 4.3 and 4.4); some classic algorithms addressing this task are presented in sections 4.5 and 4.6. The question of logical completeness, which will loom large over Part III of this book, is introduced in section 4.7. Section 4.8 discusses the computational tractability of REG algorithms, and we conclude the chapter with a discussion of the way in which salience can be given a place in the classic REG algorithms (section 4.9).

4.1 Computational Pre-history of REG

From the early days of artificial intelligence onwards, reference generation played a role in the work of practitioners. Some of them were not specifically interested in reference, but their work required the construction of REG algorithms. Let us do some archaeology and dig up their ideas.

1 The discussion of the completeness of REG algorithms in this chapter goes back to [van Deemter, 2002] and [van Deemter and Halldórsson, 2001]. The discussion of Appelt and Kronfeld's work owes a debt to [Krahmer and Van Deemter, 2012]. Most other parts of this chapter, including the analysis of the classic REG task, are new.

Referring in a world of blocks: Winograd (1972). The book "Understanding Natural Language" did much to put Computational Linguistics on the intellectual map [Winograd, 1972]. The program described in the book aimed primarily at offering a level of natural language *understanding* but performed language generation as well, as exemplified by the following dialogue fragment[2]:

Person: *What does the box contain?*
Computer: *The blue pyramid and the blue box.*
Person: *What is the pyramid supported by?*
Computer: *The box.*
Person: *How many blocks are not in the box?*
Computer: *Four of them.*
Person: *Is at least one of them narrower than the one which I told you to pick up?*
Computer: *Yes, the red cube.* (...)
Person: *Can a pyramid be supported by a block?*
Computer: *Yes* (...)
Person: *Stack up two pyramids.*
Computer: *I can't.*
Person: *Will you please stack up both of the red blocks and either a green cube or a pyramid?*
Computer: *OK.* (does it)
Person: *Which cube is sitting on the table?*
Computer: *The large green one which supports the red pyramid.* (etc.)

On a good day, SHRDLU was able to display remarkable feats of language understanding, but it used a relatively primitive REG algorithm. Where a full definite reference was required, the algorithm worked roughly as follows[3]: If the target referent's "type" makes it unique in the domain under discussion, then a noun is used that describes this type (e.g., "the box"). If an object had previously been assigned a proper name (as in "Call the biggest block 'Superblock'"), then the name would be used. Otherwise, the colour would be used ("the red cube") and, if this was insufficient to individuate the target referent, the size of the object as well (e.g., "the large green one"). If type, colour and size still did not individuate the referent, then SHRDLU checked whether the object physically supported one or more other objects, in which case these were all mentioned (e.g., "which supports a red pyramid, and ..."). If this wasn't enough, the program added a postmodifier saying "which is to the right of ...",

2 The example is taken from `http://hci.stanford.edu/winograd/shrdlu/`. The same questions, concerning the same domain, also feature in [Winograd, 1972], section 1.3.

3 For source code of SHRDLU's generation component, see `http://hci.stanford.edu-/winograd/shrdlu/code/newans`, especially the function NAMEOBJ. See also [Winograd, 1972], section 8.3.3, Naming Objects and Events.

naming all the objects located to its left. If the resulting description still did not characterize the referent uniquely, SHRDLU would generate a definite NP nonetheless.

SHRDLU's REG program had many limitations, yet it embodied some decisions that have withstood the test of time. In essence, if we abstract away from the use of proper names and indefinite NPs, SHRDLU's behaviour can be summarized as follows: a number of attributes are considered in a fixed order. These attributes are: the TYPE of the object, its COLOUR, its SIZE, the things it SUPPORTS, and the things to which it is NEAR. Each of these attributes, except the TYPE, is only included in the description if the combined attributes included so far are unable to characterize the referent uniquely. Once an attribute is included in the description it is there to stay (and note that SHRDLU includes an attribute even if it does not contribute to singling out the referent.) Later, Douglas Appelt would adopt a more sophisticated version of these ideas. Robert Dale and Ehud Reiter were to formulate them more explicitly, modify them, and defend them systematically.

Exploiting symmetries: Davey (1974). Sometimes when we appear to refer to an object, we really refer to a more general class of objects. This happens, for example, when we say "I need that screw", but what we really need is any screw of those dimensions. In other words, sometimes, we are interested not so much in a particular object r but in the equivalence class of all objects that are similar to r in some important respect. One of the earliest RE algorithms focussed on a version of this phenomenon [Davey, 1974], [Davey, 1978]. Anthony Davey's program, reconstructed ten years later in [Ritchie, 1986] and implemented by Helen Buchanan, generated textual descriptions of games of tic-tac-toe (i.e., noughts and crosses). For example,

> *I started the game by taking a corner and you took an adjacent one. I threatened you by taking the middle of the edge opposite to the corner you had just taken and adjacent to the corner I took first but you blocked and threatened me by taking the end of my edge. I forked you by taking the centre and you blocked and threatened me by taking the end of my line. I won by completing my diagonal.*

In Davey's algorithm, equivalence classes arise from *symmetries* on the board of the game. The starting point for the reference "a corner", in the description above, for example, is one specific square, for instance, the one in the top left of the 3-by-3 grid of the game. The reason why "a corner" suffices as an RE, even

though the grid has three other corners, is that these three are game-equivalent to the one in the top left. If the same move was made in a situation where not all four corners are game-equivalent (for example, because the square $b1$, adjacent to $a1$, was occupied whereas all other squares were empty), then more information would have been required (e.g., "a corner next to the square occupied by you"). The idea of grouping referents into equivalence classes has many applications: we often say things like "he put his hand on his knee", for example, without bothering to say which hand and which knee (cf., the discussion of Ludlow's ideas in section 2.4).

NPs like "my edge" and "the end of my line" make a subtle use of the history of a game. If we disregard these and other embellishments, Davey's algorithm works by computing, for each target referent and for each target line r, which entities are game-equivalent to r, then adding attributes, such as "corner", "centre", and "line" that are true of r but false of everything not game-equivalent to r, until the equivalence class of r has been characterized uniquely. Having accomplished this, the algorithm asks whether r is the only element of its equivalence class or not, and it uses this information (lines 4-7 in Algorithm 1) to choose between a definite and an indefinite description. To clarify the working of an algorithm we shall often use an informal style of pseudo-code, as in 1. Pseudo-code can leave out many details; for instance, the pseudo-code below does not say how actual NPs were constructed (lines 5 and 7), because this is not important from our present point of view.

Algorithm 1 Davey's symmetry-aware REG algorithm

Input: A representation of a tic-tac-toe board; a target cell r on the board.
Output: An NP that singles out the equivalence class of r; the NP is definite if this class contains only one element, and indefinite otherwise.

```
1: Find a description D such that:
2:    D is true of r and
3:    D is false of all items outside equivalence-class(r)
4: if D is only true of r then
5:    make a definite NP based on r
6: else
7:    make an indefinite NP based on r
```

REG programs designed in the early 1970s were clever, but they were not intended to address the main theoretical issues surrounding reference. This was to change in the next decade.

4.2 The California School

The computational generation of referring expressions was studied in the context of such dialogue systems as HAM-ANS, which were able to conduct limited conversations, which were nonetheless sophisticated in their ability to prevent the user from misunderstanding the utterances generated by the system (e.g., [Jameson, 1983], [Wahlster and Kobsa, 1989]; cf., our chapter 12). We will here focus on a group of researchers who used algorithms to explore issues that had been studied more informally by philosophers and linguists, including many of the questions discussed in chapter 2. Following [Krahmer and Van Deemter, 2012], I call this group of researchers the California School. A number of computer programs resulted from their work, and the principles underlying these programs live on.

Computational Linguistics research practices in the 1980s differ considerably from what they are at the moment. The California School views reference as a speech act formalisable in terms of the framework that models human communication using a combination of Epistemic Logic and computational planning [Cohen and Levesque, 1985]. Doug Appelt (personal communication) described the intellectual climate very well: *"(...) the research themes that originally motivated our work on generation were the outgrowth of the methodology in both linguistics and computational linguistics at the time that research progress was best made by investigating hard, anomalous cases that pose difficulties for conventional accounts."* In the Preface I suggested that reference may be seen as the fruit fly of language (see also section 16.3). By analogy, the research Appelt was referring to may be likened to studying "black swans" (cf., [Taleb, 2010]), that is, phenomena that fascinate because of their rarity. The advantage of this approach is that it zooms in on interesting phenomena from the start; as we shall see, there are disadvantages as well.

In Apple's KAMP system, speech acts are generated as part of a planning system. Suppose a computer is given the goal of repairing a machine. To meet this goal, it generates a plan in which a person removes a pump from a platform. To meet this sub-goal, the person has to be told to use a particular wrench. To allow the hearer to find it, the system offers as much information about the wrench as is required, saying for example that it is a wrench and that it is located in the toolbox [Appelt, 1985a]. Concepts and words are at the heart of this approach, but the system might also decide to point.

Choosing properties. Appelt and Kronfeld were interested in high-level questions concerning the nature of REs and their role in communication. Yet they must have thought hard about strategies for individuating a referent. The referent will tend to have many different properties; which ones should be included in the RE? Appelt and Kronfeld's starting point in this matter are the Gricean Maxims (see section 2.8). Appelt observed that the Maxims militates against overly elaborate REs [Appelt, 1985a]. Although he outlines an algorithm that is guaranteed to choose the shortest description always, he ends up arguing for a more relaxed interpretation. Using the word "descriptors" instead of our word "properties", he writes:

> *"*KAMP *chooses a set of basic descriptors when planning a describe action to minimize both the number of descriptors chosen, and the amount of effort required to plan the description. Choosing a provably minimal description requires an inordinate amount of effort and contributes nothing to the success of the action.* KAMP *chooses a set of descriptors by first choosing a basic category descriptor (see Rosch 1978) for the intended concept, and then adding descriptors from those facts about the object that are mutually known by the speaker and the hearer, subject to the constraint that they are all linguistically realisable in the current* NP*, until the concept has been uniquely identified. (. . .) Some psychological evidence suggests the validity of the minimal description strategy; however, one does not have to examine very many dialogues to find counter-examples to the hypothesis that people always produce minimal descriptions."* [Appelt, 1985a] (p. 21)

This tantalizingly brief description contains three themes that later research has elaborated on: the non-minimality of human-produced REs, the role of Rosch-style basic categories [Rosch, 1978], and the idea that properties may be added one by one until the referent has been identified. Computational linguists have used the word "incremental" to characterize this procedure. All these themes will be revisited when we discuss later algorithms.

Adding information. So far, we have been talking as if it were the sole goal of an RE to single out a referent. Appelt and Kronfeld recognized, however, that reference can serve several purposes and sought to explain how this can happen. In particular, they showed how an RE can add information about a referent, saying things about it that the recipient does not yet know. Using the perspective proposed in section 1.7, they demonstrated computationally how Information Sharing can work in two directions at the same time, namely,

from shared information to the intended referent (the normal direction), and from the intended referent to shared information (the direction discussed by Barwise and Perry). By pointing to a tool, for example, we can identify the tool. If, in addition to pointing, we also say "the wheelpuller", then the descriptive content of the RE may serve to inform the hearer about the *function* of the tool.

To see how an RE can add information, consider the reference to a wheelpuller r. The action by the speaker S of describing r to the hearer H using a conjunction of properties $P_1 \wedge ... \wedge P_n$, or simply by pointing, is subject to a set of preconditions (which we simplify slightly here):

(1) S believes $P_1(r) \wedge ... \wedge P_n(r)$
(2) $\neg$ (H believes $\neg(P_1(r) \wedge ... \wedge P_n(r))$)
(3) $\forall x(\neg(H \text{ believes } \neg(P_1(x) \wedge ... \wedge P_n(x))) \rightarrow x = r)$

In plain English, (1) the speaker believes that the properties expressed are true of r, (2) the hearer does not believe the conjunction of properties to be false of r, and (3) r is the only object x such that the hearer doesn't believe the conjunction of properties to be false of x. Note the subtle use of double negation: the hearer need not believe that all of the properties in the description are *true* of r, as long as she doesn't believe that any of them are *false* of r. This means that the preconditions for referring to r as "the wheelpuller", while also pointing at it, are fulfilled as long as the speaker believes r to be a wheelpuller and the hearer doesn't have reasons to believe it isn't. It is this double negation that allows referring speech acts to smuggle in information that is not yet in mutual knowledge (cf., [Appelt and Kronfeld, 1987]). Epistemic embeddings in the style of Clark and Marshall (section 3.1) are not considered.

There is something frustrating about research papers from the California School (cf., [Krahmer and Van Deemter, 2012]). On the one hand, they contain genuine insights into the complexities of communication. On the other hand, it is remarkably difficult to find out how the programs described actually worked, since code was lost and much of what was written about it is pitched at a high level of abstraction. Most important of all, these insights were not tested empirically. As we shall see in the next chapter, this style of work has gone out of fashion, at least in Computational Linguistics, giving way to a skepticism about methods that require a "deep" modelling of a speaker's or hearer's knowledge. We shall also see, however, that many of these authors' basic ideas have survived. Moreover, it is worth realizing that these researchers had to code all deductive mechanisms from scratch. Given

the substantial progress in computational theorem proving in the last decades (e.g., [Robinson and Voronkov, 2001]), the time may have come to re-assess these ideas. The investigations in our chapters 10 and 11, where we look at logic and theorem proving, are a step in this direction.

Relevance. Kronfeld saw that different utterance situations favour different descriptions. He gives the example of the sentence "The city with the world's largest Jewish community needs more policemen", whose subject NP is intended to refer to New York [Kronfeld, 1989]. This NP is conversationally irrelevant, unless the speaker wants to suggest that the ethnic make-up of the city explains the need for more police. Even though no computational mechanism was offered for avoiding irrelevant REs, Kronfeld's insistence on conversational relevance is well taken. Kronfeld's challenge has seldom been discussed in the literature on REG, yet I will argue that some promising inroads into this difficult area have been made (see e.g., section 16.2). In fact, we shall see that some aspects of this issue lend themselves well to statistical methods. For example, the preceding context of an RE may be used (e.g., employing n-grams or a more semantics-aware mechanism) to rate some properties as more relevant than others and to favour their selection in REG.

4.3 The Classic REG Task

In recent years, the California School has been largely superseded by a new research tradition. The issues studied by Appelt and Kronfeld are still discussed from time to time [Heeman and Hirst, 1995, Stone and Webber, 1998, O'Donnell et al., 1998, Koller and Stone, 2007]. A different approach gained prominence, however, not because the earlier approach had been falsified, but rather because it was so ambitious that it was difficult to falsify. Two authors played a particularly important role in this shift, namely, Robert Dale and Ehud Reiter. Somewhere around 1990, these authors started re-focussing on the smaller problem of determining what properties an RE should use if identification of the referent is the central goal [Dale, 1989a, Reiter, 1990a]. This line of work culminated in the seminal article *Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions* [Dale and Reiter, 1995]; in this work, other aspects than identification were temporarily disregarded, not because they were deemed unimportant, but in order to concentrate on simple things first.

We shall examine the algorithms that came out of this period of re-focussing, paying particular attention to their underlying assumptions. It will be helpful to use some terminology introduced by philosophers of language (section 2.3) calling the set of elements that share a property P the *denotation* or *extension* of P, abbreviated $[\![P]\!]$.

> Given is a communicative situation involving a finite domain M of entities and an element $r \epsilon M$, which is the target referent. Given are also one or more other elements of M, the distractors; furthermore, a finite set $\mathcal{P}$ of atomic properties (i.e., properties without logical structure, or whose logical structure is ignored), each of which holds true of one or more elements of M. Thus, for every property $P \epsilon\ \mathcal{P}$, the extension $[\![P]\!]$ of this property is a non-empty subset of D.

Unless stated otherwise, the remainder of this section will use the term "properties" as denoting *atomic* properties. To see how the reference generation task may be understood, continuing to focus on Content Determination, let us first examine its most straightforward version. The well-known correspondence between set-theoretical operations like *set union* (which applies to extensions of properties) and propositional logic connectives like *disjunction* (which applies to propositions) will allow us to mix the two terminologies.[4] Thus, I shall freely speak about "conjoining" or "intersecting" properties depending on which of the two perspectives suits me.

Consider the following perspective on REG:

> *The classic* REG *Task (naive version).* If there exists a set of properties $\{P_1, .., P_n\}$, where each P_i is an element of $\mathcal{P}$, and where the conjunction of all the P_i in the set singles out the referent r (i.e., $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$), then find such a set and conjoin its elements. If no such set of properties exists, then say so.

This task, however, is suspiciously easy. A simple solution would be to collect all the properties that are true of the referent r, as in Algorithm 2 below. In practice, researchers do not focus on what we called the naive REG task; they employ an additional constraint that is often left implicit. As we saw in section 1.5, sometimes the (somewhat vague) idea is to produce a "felicitous" RE but, increasingly, researchers are becoming more specific. For example, they may

4 When we speak about "conjoining" properties P and Q we mean, logically speaking, conjoining the atomic properties $\lambda x.P(x)$ and $\lambda x.Q(x)$, forming the complex property $\lambda x.(P(x) \wedge Q(x))$.

Algorithm 2 The Total Reference Algorithm for conjunctive REs

Input: A domain of objects, containing a target referent r and a non-empty set of distractors. A set $\mathcal{P}$ of properties defined over the domain.
Output: A distinguishing description $\mathcal{D}$ of r that uses conjunctions of properties in $\mathcal{P}$, if such a distinguishing description exists.

1: Let $\mathcal{P}'$ be the set of all properties in $\mathcal{P}$ that are true of r
2: Let $\mathcal{D}$ be the conjunction of all elements of $\mathcal{P}'$
3: **if** $[\![\mathcal{D}]\!] = \{r\}$ **then**
4: **return** $\mathcal{D}$
5: **else**
6: **return** "No distinguishing description of r exists"

be aiming to generate RE that have optimal utility for hearers; more often, the aim is human-likeness:

> *The classic* REG *task (main version).* If there exists a set of properties $\{P_1, .., P_n\}$, where $P_i \in \mathcal{P}$, and where the conjunction of all the P_i in the set singles out the referent r (i.e., $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$), then find such a set and conjoin its elements. If no such set of properties exists, then say so. Furthermore, make sure that $\{P_1, .., P_n\}$ are collectively as similar as possible to the set of properties that human speakers would use if they were to refer to r in the situation at hand.

This task definition[5] is in line with a view of REG as mimicking human behaviour. Note that verifying the second requirement cannot be done without hard (e.g., experimental) graft; the same would be true if the task definition focussed on properties that are of optimal benefit to hearers (cf., section 1.5).

To see what's at stake, consider a well known problem in Artificial Intelligence, namely, computer chess. It's challenging to make a program that plays chess strongly; it's much harder to make a program that also plays in a human-like style: to do that, difficult questions need to be answered: What strategies are preferred by human players (e.g., when two moves are equally strong)? Moreover, playing "like a human" is, pending further specification, not a well-defined goal because people differ in terms of their abilities and inclinations (e.g., a disposition towards defence or attack). Issues of this nature will occupy

5 From now on, we shall often omit the part of an algorithm that specifies what happens if no distinguishing description is found (e.g., lines 5 and 6 of Algorithm 2). This will allow us to focus on the core of the algorithm.

us in the next chapter, where we shall be concerned with the empirical validation of REG algorithms. Finally, the task definition could be even more radically empirical, using human-likeness as its only criterion, regardless of whether human speakers manage to single out the referent:

> *The classic* REG *task (radically empirical version).* Find properties $\{P_1, .., P_n\}$ in $\mathcal{P}$ that are collectively as similar as possible to the set of properties that human speakers would use if there were to refer to r in the situation at hand.

In the remainder of this chapter and the next, we shall focus on the second task definition, but there will be echoes of the third one in Part IV of the book.

4.4 Assumptions Behind the Classic REG Task

Before discussing algorithms that address the classic REG task, it will be useful to state some ideas that are implicit in this task. Some are best seen as simplifying assumptions, whereas others are genuine preconceptions about the goal of REG (henceforth, presuppositions). In what follows, we will make these ideas explicit. Parts III and IV of this book will explore what happens when some of these assumptions and presuppositions are abandoned. Readers who are eager to see the classic REG algorithms could jump to section 4.5 directly.

Assumption 1: Reference is always to a single individual. The task definition presupposes that the target of an RE is always just one object, not a larger set. As long as this assumption is made, plural NPs (e.g., "the black poodles", "cats and dogs") are not generated. As we shall see in chapter 8, reference to sets turns out to be more difficult than reference to singular objects, giving rise to many new research questions.

Assumption 2: All domain objects are equally salient. Things present themselves to us with different degrees of urgency. Suppose you and an academic colleague are talking. Other things being equal, the people in your department are more salient than the ones in other parts of the School; these in turn are more salient than most other people. All these people are referable in principle, but the more salient ones take less verbal effort. For example, to refer to your own Head of Department, you might say "the Head", but in other cases the name of the department needs to be stated (e.g., "the Head of Maths"). The classic REG task definition says nothing about differences in salience. See section 9.7 for discussion.

Assumption 3: Context-dependent and vague properties are not used. It is evident from the classic task definition that a property must either apply or not apply to a given domain object. Properties must be crisp and well defined, without depending on context. Yet vague properties (like "old" or "large") are notoriously context-dependent, and so are crisp properties like "leftmost": a book on a shelf may be leftmost when only the volumes of an encyclopaedia are considered, but not if the entire shelf is taken into account. Chapter 9 will address these issues.

Assumption 4: REs do not express complex properties. The properties accumulated by classic REG algorithms are not themselves composed of logically simpler properties, as when we say "the cow that is *not* brown". Thus, in Dale and Reiter's work, conjunction is the only logical operation permitted, to the exclusion of operations like negation.

Assumption 5: REs do not express relations. The task definition precludes the use of relations. We shall see in section 6.4 that some types of REs that make use of 2-place relations (e.g., "the cup on the table") are nonetheless generated by some early algorithms (section 6.4) and by the approach discussed in section 6.5. More complex relational descriptions are discussed in chapter 10. REs that use relations with more than 2 argument places ("the present given by Joey to Marc") have yet to be addressed.

Assumption 6: Sets are finite. Reference is not limited to finite domains. It is possible, for example, to refer to a natural number (e.g., "the only even prime number"), using all other numbers as distractors. In practice, however, REG has focussed on finite M, $\mathcal{P}$, and L, as we have seen. Thus, in practice, REG algorithms deliver a finite set of properties $\{P_1, .., P_n\}$ whose intersection denotes the referent. Infinite sets are discussed briefly in section 4.7 and chapter 10.

Assumption 7: Content Determination precedes realization. In most REG algorithms, the semantic content of the description is chosen at the outset, with other NLG tasks – such as Lexical Choice and Linguistic Realization – only starting once Content Determination is finished. This setup only makes sense if there exists a suitable Linguistic Realization for every property. Problems arise if no words can be found for expressing a given concept, causing what is known as a *generation gap* [Meteer, 1991]. This assumption is shared by most algorithms; for exceptions see section 6.6.

In Part III, where extensions of the classic REG problem are discussed, we shall see what happens when some of these assumptions are abandoned. At a

more fundamental level, the REG task relies on the following presuppositions, relating to the type of *Information Sharing* that the speaker engages in:

Presupposition 1: REs always identify the target referent. In Part IV of the book, we shall encounter situations where precise identification is not feasible, so at best only approximate descriptions can be produced, which (for example) fail to exclude certain distractors. Perhaps the clearest examples arise when speakers verbally describe a geographical region, which can often only be done approximately. Approximate descriptions (discussed in chapter 14) are precluded by the present task definition.

Presupposition 2: Identification is the only goal of reference. This is implicit in the REG task definition, yet various authors have argued that REs can serve other functions as well. Dale and Reiter offered the example "Don't sit at the newly painted table", where the description of the table may not only serve to identify the referent, but to explain the reason for the advice. To let REG algorithms focus on identification alone is an abstraction of reality: reference serves other purposes as well. We shall turn to these issues in chapter 15.

Perhaps the trickiest presupposition underlying Dale and Reiter's framing of the reference task pertains to its wider context. The crucial concept in this connection is the concept of common knowledge, which we discussed in chapter 3. After all, why compose a set of properties if you cannot be sure that the hearer shares your understanding of these properties, in terms of what their extensions are? We distinguish two aspects of this issue.

Presupposition 3: $\mathcal{P}$ represents speaker and hearer's common ground. It is widely thought that a property can only enter an RE if its meaning is in the speaker and hearer's common ground. Details have typically been less clear; Dale and Reiter, for example, offered no computational mechanism for distinguishing information in common ground from other information. In chapter 13, we shall argue that common ground is frequently problematic and investigate what this means for REG.

Presupposition 4: The extensions of the properties in L are obvious. The task definition is limited to situations where hearers have direct access to the extensions of all the properties in L. In chapter 12, we discuss situations in which hearers need to work hard to discover the extensions of some crucial properties of the referent, such as its location in a room.

Part IV investigates what happens when these Presuppositions are abandoned.

4.5 Exploring the Gricean Angle Computationally

Having proposed a re-framing of the REG problem, Dale and Reiter went on to explore a range of REG algorithms. In doing so, they used many of the same ideas as their predecessors, starting from the assumption (also invoked by Doug Appelt) that the content of an RE is best understood in terms of the Gricean Maxims [Grice, 1975], whose Maxim of Quantity we re-state here:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

This Maxim came to be associated with an approach known as Full Brevity (FB). This approach, discussed in [Appelt, 1985a] and formalized in [Dale, 1989a], says that the number of properties in the Logical Form need to be minimized. Dale proposed the mechanism described in Algorithm 3 for enforcing FB. The first step checks whether there is a single property of the target that rules out all distractors. The second step checks whether any combination of two properties does this, and so on, until the referent has been singled out or until conjunctions of all possible lengths have been attempted. Note that this does not tell us what Logical Form is generated if, for example, step (1) can choose between two distinguishing descriptions of equal length. To turn the procedure into a complete algorithm, a tie-breaking rule would need to be added.

Algorithm 3 Dale's Full Brevity (FB) algorithm

Input: A domain of objects containing a target referent r and a non-empty set of distractors. A set $\mathcal{P}$ of n properties true of r.
Output: A distinguishing description $\mathcal{D}$ of r that uses conjunctions of properties in $\mathcal{P}$, if such a distinguishing description exists. If $\mathcal{D}$ is found, then no purely conjunctive distinguishing description of r exists that uses fewer properties.

```
Look for a description that distinguishes r using one property
if a description D is found then
    return D
else
    Look for a description that distinguishes r using two properties
    if a description D is found then
        return D (else Etcetera, using up to n properties)
```

It is possible to believe that brevity is the key factor in the quality of an RE, but to take it with a pinch of salt. One way to do this would be to seek inspiration from information theory [Shannon, 1948]: select properties one by one, always choosing a property that divides the set of distractors as evenly as possible between those individuals that have the property and those that do not. If it's always possible to find a perfectly "even" property (i.e., one that is true for exactly 50% of the remaining distractors), then a mere n properties distinguish 2^n different individuals, hence 20 properties would suffice to nail down any individual within a domain of 2^{20} individuals.[6] One could use this idea in REG, always choosing properties that can be expected to do well. But it is possible to do better, because the speaker knows who she intends to refer to. So, rather than selecting a property that is expected to do best on average, the speaker can select the property that is certain to do best for this particular referent. The name of this algorithm is the Greedy Algorithm.

The Greedy Algorithm ([Dale, 1989a], henceforth GR) selects properties one by one, always choosing the property that is true of the intended referent and excludes the greatest number of distractors. GR does not always produce a minimal description, because a property that removes the maximum number of distractors at the time of its inclusion might not remove the maximum number of objects in combination with properties that will later be added (because these are not known yet). Recall that the REG Task Definition requires the domain to contain the referent, and one or more other entities as well. The algorithm starts by initializing the description $\mathcal{D}$ to be empty. At the core lies step 3, where a property is selected whose Discriminatory Power (DP) is second to none: it removes the maximum number of distractors, in other words. Once again, the literature does not specify how to choose if there is a tie (i.e., when several properties remove the same number of distractors).

In the Greedy Algorithm, properties are added to the description one by one, and never retracted. Since this will prove to be a common feature of many algorithms, I shall give it a name: the *monotonic* approach to REG.[7]

6 This idea brings to mind the parlour game known as *Twenty Questions*, where one player has to ask a series of yes/no questions to identify, as quickly as possible, an individual the other player has in mind: "A politician?" Yes. "A Democrat?" No. "Been a President?" Yes. *Etc.*

7 The term *monotonic* REG, an allusion to *(non)monotonic* logic, was chosen because the term *incremental* REG (which would have been apt in some ways) has come to be associated with one particular type of monotonic approach, as we shall see.

Algorithm 4 The Greedy (GR) Algorithm

Input: A domain of objects containing a target referent r and a non-empty set of distractors M. A set $\mathcal{P}$ of properties true of r.
Output: A distinguishing description $\mathcal{D}$ of r that uses conjunctions of properties in $\mathcal{P}$, if such a distinguishing description exists.

```
1: Start out with an empty D
2: while Not all distractors have been ruled out and P ≠ empty do
3:     Select a new property P from P, choosing one whose DP is maximal
4:     if P is false of some distractors then
5:         Add P to D
6:         Remove P from P
7:         Remove from M all distractors ruled out by P
```

Using a more formal set notation, the algorithm can be rendered as follows:

```
1: D := φ
2: while M ≠ ∅ do
3:     Select a new property P from P, (...)
4:     if M ⊈ [[P]] then
5:         D := D ∪{P}
6:         P := P −{P}
7:         M := M ∩ [[P]]
```

I have used dots to replace the basis on which a new property P is chosen, revealing an algorithmic "skeleton" that the Greedy Algorithm shares with other algorithms. We shall soon encounter another way in which the dots may be filled. For ease of reading, however, we shall stick with the informal style displayed in algorithm 4, always assuming that phrases like "Add so-and-so to $\mathcal{D}$" are interpreted in accordance with the set-theoretic explication above.

To see how these algorithms behave, let's look at a specially constructed domain. We separate each property into an attribute (like COLOUR) and a value (like RED), grouping properties into clusters. Suppose our domain contains the dogs $\{a, b, c, d, e, f, g\}$. We use five different attributes, each of which has two values, denoted as val_1 and val_2. We choose values that are each other's complement, so the choice between them is obvious once the referent is given:

HAIRINESS $val_1 = \{c, e, f\}$ (HAIRY), $val_2 = \{a, b, d, g\}$ (SHORT-HAIRED)
COLOUR $val_1 = \{a, b, c, e\}$ (BLACK), $val_2 = \{d, f, g\}$ (WHITE)
CLASS $val_1 = \{a, b, e, f\}$ (MONGREL), $val_2 = \{c, d, g\}$ (PUREBRED)
HARDINESS $val_1 = \{d, e, f, g\}$ (OUTDOOR), $val_2 = \{a, b, c\}$ (INDOOR)
TAIL-STATUS $val_1 = \{a, b, c, d, e, f\}$ (WITH TAIL), $val_2 = \{g\}$ (WITHOUT TAIL)

Consider the target referent a. The easiest way to see that a cannot be singled out using the ten properties at hand is by pondering the Total Reference algorithm of section 4.3. This algorithm would conjoin all properties true of a, generating $\{a, b, d, g\}$ (HAIRY) $\cap$ $\{a, b, c, e\}$ (BLACK) $\cap$ $\{a, b, e, f\}$ (MONGREL) $\cap$ $\{a, b, c\}$ (INDOOR) $\cap$ $\{a, b, c, d, e, f\}$ (WITH TAIL), and this intersection equals the set $\{a, b\}$. No intersection of properties that includes a can be smaller, so a cannot be individuated by conjoining atomic properties. The reason is that every property whose extension includes a includes b as well: as we shall say, b is a *satellite* of a. Similarly, b cannot be individuated, because a is a satellite of b. Satellites [van Deemter and Halldórsson, 2001] will take on a life of their own in chapter 10.

To see how FB and GR can produce different outcomes in this situations, consider the target referent e, starting with the Full Brevity algorithm. FB will combine $\langle$COLOUR: $val_1\rangle$ (BLACK, which is true of the dogs $\{a, b, c, e\}$ only) with $\langle$HARDINESS: $val_1\rangle$ (OUTDOOR, which is true of $\{d, e, f, g\}$), because these two properties jointly single out e while all other Logical Forms that manage the same feat happen to contain three or more properties. The output thus consists of the properties BLACK and OUTDOOR, which may be realized as "the black outdoor dog", "the only black dog", and so on. We are adding the word "dog" to fill the position of the noun, even though it does not contribute to distinguishing the referent. We shall return to this issue later, when the TYPE attribute is discussed.

Although GR often produces the same Logical Form as FB, this is not the case in the present example. GR will start by selecting the property $\langle$HAIRINESS: $val_1\rangle$ (HAIRY, corresponding to the set $\{c, e, f\}$), because it is the property that excludes the most distractors. Even though only two distractors are left, namely, c and f, no single property manages to remove both of these. As the next step, GR will select either $\langle$COLOUR: $val_1\rangle$ (BLACK, removing f) or $\langle$CLASS: $val_1\rangle$ (MONGREL, removing c), but in each case a third property is required to remove the last remaining distractor. The example shows that GR leaves some issues undecided: it allows e to be described using either the Logical Form {HAIRY, BLACK, OUTDOOR} or the Logical Form {HAIRY, BLACK, MONGREL}, for example. The FB algorithm doesn't always decide either: the individual c, for example, can either be described minimally using {INDOOR, PUREBRED} or using {HAIRY, PUREBRED}.

4.6 The Incremental Algorithm

It is time to introduce the algorithm proposed by Dale and Reiter, which has come to be known as the Incremental Algorithm (IA). Around the end of the last century, Helmut Horacek wrote that "the incremental algorithm is generally considered best now" [Horacek, 1997]; Jon Oberlander added that it "clearly achieves reasonable output much of the time" [Oberlander, 1998]; and Emiel Krahmer wrote "the Incremental algorithm has become more or less accepted as the state of the art for generating descriptions" [Krahmer and Theune, 2002]. It is still the best-known REG algorithm, and our own experiments (e.g., chapter 5) confirm that it is doing something right, although we shall also raise problems, which will ultimately cause us to regard it with reservation (see 16.2; and from a viewpoint of understanding variation in language production, section 6.3).

We have seen in chapter 3 that some properties are more "preferred" than others, and that these differences in preference make themselves felt in the REs that human speakers produce. Though difficult questions can be asked about the reasons behind these differences, we have seen that their effects can be modelled by listing properties in order of their importance. Given such a list, the algorithm examines its properties one by one, starting with the most "preferred" one. A property is added to the Logical Form if the property helps with the identification of the referent (i.e., if it removes one or more distractor objects). The algorithm halts when no more progress can be made (i.e., when none of the properties in the list has anything further to contribute to the identification of the referent). This, in a nutshell, is the Incremental Algorithm; details and examples will be offered presently.

Before we proceed, it should be noted that the order in which properties are examined does not necessarily have any bearing on the order in which they are realized linguistically into a noun phrase (cf., sections 1.4 and 9.5). Note, furthermore, that not too much should be read into the word "preferred": to regard the first property in the list (also known as a Preference Order) as the most preferred one is little more than a manner of speaking (cf., section 3.4).

Dale and Reiter offered two arguments for the IA. The first is of a purely computational nature: it argues that the problem of finding a Logical Form that contains the *minimum* number of properties, as required by the Full Brevity algorithm, is computationally intractable; we shall later devote a separate section (section 4.8) to this argument, which history appears to have almost

forgotten. The second argument, which has attracted more attention, is the well-known fact that people often produce non-minimal descriptions (e.g., [Pechmann, 1989]). These ideas were rooted in the earlier literature on the subject, as we have seen (e.g., [Appelt, 1985a]).

It is worth comparing the idea of the Incremental Algorithm to the Greedy Algorithm, with which it has much in common: both algorithms operate monotonically, adding properties one by one until the referent has been singled out, without ever withdrawing a decision. The difference is that GR decides on the basis of Discriminatory Power, whereas IA decides on the basis of Intrinsic Preference (cf., section 3.5). Thus, the Greedy Algorithm (GR) is entirely motivated by logical considerations to do with the extension of properties, whereas IA is blind to the extension of a property. Using the terminology of chapter 2, this makes GR extensional and IA intensional. Suppose, for example, the things that have lungs are the same as those that have kidneys; this does not prevent IA from distinguishing between the properties "has lungs" and "has kidneys". The flip side of this situation is that GR can be seen as having greater explanatory power, because it offers a *reason* why one property is chosen instead of another. The IA is unsystematic by comparison, unless it can offer a reason why one property is preferred over another.

Because IA is structurally so similar to GR, it can be presented in much the same way. We can use the same pseudo-code, except that the choice in line 3 is guided not by Discriminatory Power (which varies during the course of the Greedy Algorithm's execution) but by a fixed Preference Order.

Algorithm 5 IA_{Prop}: the Incremental Algorithm based on properties

Input: A domain of objects, containing a target referent r and a non-empty set M of distractors. A set $\mathcal{P}$ of properties of r. A linear Preference Order defined on $\mathcal{P}$.
Output: A distinguishing description $\mathcal{D}$ of r that uses conjunctions of properties in $\mathcal{P}$, if such a distinguishing description exists.

1: Start out with an empty $\mathcal{D}$
2: **while** Not all distractors have been ruled out and $\mathcal{P} \neq$ empty **do**
3: Select a new property P from $\mathcal{P}$, *choosing the most preferred one*
4: **if** P is false of some distractors **then**
5: Add P to $\mathcal{D}$
6: Remove P from $\mathcal{P}$
7: Remove from M all distractors ruled out by P

The algorithm finds better and better approximations of the target set $\{r\}$ by adding more and more properties to $\mathcal{D}$. There is no backtracking, so if some

property P_i in $\mathcal{D}$ is made redundant by later additions, then P_i is retained as a member of $\mathcal{D}$ nevertheless.

However, what Dale and Reiter proposed is not quite IA_{Prop}, but a slightly more elaborate algorithm, which separates attributes and values, and which we shall call IA_{Att}. The main intuition behind IA_{Att} is that attributes permit a natural way of grouping similar properties together, making it easy to generalize over them, for example, by saying that all properties that express the COLOUR attribute occur earlier in the Preference Order than any other attributes. Furthermore, the use of attributes allows a closer approximation to GR, choosing between the different values of a given attribute on the basis of (mainly) Discriminatory Power.

IA_{Att} is presented schematically in Algorithm 6 (below). This time it is attributes, not properties, that are listed in a Preference Order $\mathbf{A}$. If A_i precedes A_j in $\mathbf{A}$, then A_i is preferred over A_j; as a consequence, A_i will be considered before A_j by the algorithm. Given an attribute, `FindBestValue` selects the value that removes most distractors while still including the target r. In case of a tie (i.e., no value removes more distractors than all others), `FindBestValue` chooses the least specific of the contestants. For example, in a situation where the property of being a dog rules out as many distractors as being a chihuahua, the latter cannot be chosen.

$\mathbf{A}$ is the list of attributes; L is the set of attribute/value combinations returned by the algorithm. A further notational convention will be useful: values will be identified by two indices, the first of which identifies the attribute. Thus, to denote value j of attribute A_i, we write $V_{i,j}$. This version of the algorithm, which is schematized in Algorithm 6, will be called IA_{Att}. The initialization of L is omitted for brevity. Note, once again, the restriction to distinguishing descriptions that are conjunctions of properties in $\mathcal{P}$; in the remainder of this book, we shall omit this restriction to keep the presentation of algorithms readable; the issue will be discussed fully in chapter 8.

Following [Appelt, 1985a], Dale and Reiter added a provision to ensure that each Logical Form generated contains a TYPE: if no TYPE is selected during the normal course of their algorithm, it is added to the Logical Form at the end. This treatment ensures that every Logical Form contains one property realisable as a noun. We shall accept this with one qualification: to prevent unnaturally lengthy descriptions, we *also* assume that the TYPE attribute is placed at the head of the Preference Order, causing it to be considered before all other attributes. Without this qualification (i.e., if the TYPE is always only

Algorithm 6 IA_{Att}: the Incremental Algorithm based on attributes

Input: A domain of objects, containing a target referent r and a non-empty set M of distractors. A set $\mathcal{A}$ of attributes at least one of whose values is true of r. A linear Preference Order defined on $\mathcal{A}$.
Output: A distinguishing description $\mathcal{D}$ of r that uses conjunctions of properties in $\mathcal{P}$, if such a distinguishing description exists.

1: Start out with an empty $\mathcal{D}$
2: **while** Not all distractors have been ruled out and $\mathcal{A} \neq$ empty **do**
3: Select a new attribute A_i from $\mathcal{A}$, *choosing the most preferred one*
4: $V_{i,j} :=$ `FindBestValue`(r, A)
5: **if** $V_{i,j}$ is false of some distractors **then**
6: Add $V_{i,j}$ to $\mathcal{D}$
7: Remove A from $\mathcal{A}$
8: Remove from M all distractors ruled out by $V_{i,j}$

added as an afterthought), the IA can easily include properties in a Logical Form that are made superfluous by the TYPE. As explained in chapter 5, types can be treated similarly in all REG algorithms. At the end of the book, in section 16.3, we shall turn to the question of what types really are.

Let's see how the IA behaves when applied to the example that we used in order to exemplify FB and GR. We repeat the example for convenience.

HAIRINESS $val_1 = \{c, e, f\}$ (HAIRY), $val_2 = \{a, b, d, g\}$ (SHORT-HAIRED)
COLOUR $val_1 = \{a, b, c, e\}$ (BLACK), $val_2 = \{d, f, g\}$ (WHITE)
CLASS $val_1 = \{a, b, e, f\}$ (MONGREL), $val_2 = \{c, d, g\}$ (PUREBRED)
HARDINESS $val_1 = \{d, e, f, g\}$ (OUTDOOR), $val_2 = \{a, b, c\}$ (INDOOR)
TAIL-STATUS $val_1 = \{a, b, c, d, e, f\}$ (WITH TAIL), $val_2 = \{g\}$ (WITHOUT TAIL)

The outcome depends on the Preference Order. If the attributes are attempted in the order in which they were listed, then HAIRINESS is attempted first, followed by COLOUR, in which case a version of GR is mimicked. But if COLOUR is attempted first, followed by HARDINESS, then the same Logical Form is generated as the one produced by FB. In other cases, much lengthier Logical Forms can result, for example when TAIL-STATUS is attempted first (narrowing down the set of possible referents to $\{a, b, c, d, e, f\}$), and then COLOUR (narrowing it down to $\{a, b, c, e\}$), followed by HAIRINESS (resulting in $\{c, e\}$) and finally HARDINESS (resulting in the set $\{e\}$). The resulting Logical Form might be realized as *the hairy black dog with a tail, who sleeps outdoors*. The ability of IA to mimic other algorithms will complicate the assessment of its ability to mimic human production, as we will soon see.

Unfortunately, IA_{Att} is affected by problems that call into question the reasons for separating properties into attributes and values, and even the idea of a fixed Preference Order itself (see also sections 3.4 and 6.3). For example, IA_{Att} makes the dubious assumption that all the values of a given attribute must be preferred to the same degree. But the colour of a blue strawberry is more likely to be noted than that of a red one, because the prototypical strawberry is red; similarly, a three-legged dog will have a higher chance of having its number of legs noted than a four-legged one [Mitchell et al., 2013b]. Different considerations point in the same direction, suggesting that the different values of a given attribute may differ sharply in terms of their degree of preference. Consider attributes such as distance, weight, or price: if you bought this cadillac for only $500, wouldn't this make you more likely to mention the price than if you had paid a more normal price? Surely, the *extremity* of a property should be taken into account when determining whether the property is worth mentioning. These issues are discussed in chapter 9 and more fully in Part IV.

Chapter 5 will discuss systematic empirical evaluation of the IA. Like all other empirical studies that I am aware of, and despite first appearances, this evaluation will essentially focus on the simpler IA_{Prop}, not the better known IA_{Att}. For although the discussion will be framed in terms of attributes, the focus will be on situations in which no two values of the same attribute can apply to any given object; consequently, the algorithm is never faced with a choice between different values. Choosing an Attribute, in these situations, is equivalent to choosing a property.

Before we go there, however, it will be interesting to explore some of the formal properties of the IA. Our first question will be straightforward: Is the algorithm always able to produce a distinguishing description? More precisely, does the algorithm produce a distinguishing description whenever one exists?

4.7 Logical (In)completeness

Let's call a REG algorithm *successful* with respect to a given situation, characterized by a Knowledge Base and a given target r, if the algorithm produces a distinguishing description in that situation. We will call an algorithm *(logically) complete* if it is successful in every situation in which a distinguishing description exists. Success is not always possible: the properties in the Knowledge Base may not be sufficient for individuating a given object; such situations should not be held against an algorithm.

There is one problem with this tidy perspective: what descriptions are possible depends on what logical operations are available. The algorithms discussed in this chapter generate Logical Forms that contain logical conjunction (i.e., set intersection) as their only Boolean operation. We therefore define a REG algorithm to be *intersectively complete* if it has the following property: whenever a referent can be individuated by intersecting a finite number of properties, the algorithm will find such an intersection. We would like to prove the Incremental Algorithm to be intersectively complete, but we shall meet a few obstacles. The first obstacle arises from overlapping values, the second from infinite sets.

Overlapping values. We start our formal investigation by considering a problem for the attribute-based version of the Incremental Algorithm, IA_{Att}. One assumption without which this algorithm cannot be proven to be intersectively complete concerns the semantic relation between different values of a given attribute: their extensions should not overlap. Values can overlap for different reasons. Colours can overlap, for example: some objects may count as both red and orange. Also, values may derive from particular parts or aspects of an object; for example, an object may be listed as both metal and plastic, because it has parts made of metal and parts made of plastic. Other kinds of examples arise if the Knowledge Base models relations through unanalysed properties. For example, a desk, or a particular *type* of desk, can stand in a given relation to more than one other company. To illustrate the problems arising from overlapping values, consider the following situation, in which companies buy particular types of desks ($\{a, b, c, d, e, f\}$):

BOUGHT-BY: PHILIPS ($\{a, b, e\}$), APPLE ($\{a, c, d, f\}$)
COLOUR: BROWN ($\{a, b\}$), YELLOW ($\{c, d, f\}$)

Desks of type a were bought by two different companies, causing the values of BOUGHT-BY to overlap. The first problem arises if a is the intended referent, and BOUGHT-BY is more preferred (in terms of the IA's Preference Order) than COLOUR. The value PHILIPS (being the BestValue of BOUGHT-BY, because it removes more distractors than the value APPLE) is chosen first, reducing the initial set of desks to $\{a, b, e\}$. Now, having found a local maximum, the algorithm is doomed to end in failure, because COLOUR is unable to remove the unwanted b without also sacrificing a. None of this can be corrected, since the algorithm does not use backtracking. Note that a distinguishing description of a would have been possible if only APPLE had been chosen instead of PHILIPS, leading to a Logical Form that could be worded as "the brown desk bought by Apple".

The second problem is even more worrying. Suppose all desks have the same colour. Now the only way to refer to a is to select two different values of the same attribute, referring to it as the desk that's bought by both Philips and Apple. IA_{Att} does not allow this, because it will always find only one best value. Once again, the algorithm fails when success is perfectly achievable.

How can these problems be remedied? If logical completeness is to be safeguarded, then what the algorithm needs to find is not the single best value for a given attribute, but the best *combination* of values: in the example above, selecting the values PHILIPS and APPLE is better than selecting either of them; in situations where other combinations of values are possible, these combinations can be compared in the manner proposed by Dale and Reiter. Applied to our original example, the revised algorithm would produce "the desk bought by Apple *and* by Philips". Such descriptions appear to be quite natural. In fact, it seems that identifying a simply as being bought by Philips can give rise to the false implicature that a was *not* bought by Apple. This suggests that the proposed algorithm might also be linguistically on the right track. Needless to say, this hypothesis would need to be investigated empirically. Remarkably, however, no empirical investigations into this area of the Incremental Algorithm have been done: Dale and Reiter's proposal for choosing between the different values of an attribute has been effectively ignored. Be that as it may, the algorithmic problems with IA_{Att} will appear in a different light when we discuss the virtues of incrementality (section 16.2), where doubts will be cast on the central idea of grouping together properties within attributes.

Infinity. To prove intersective completeness, some cardinality assumptions need to be made. These assumptions are unlikely to cause problems for applied NLG, but might have some interest if we aim to understand reference in general. For example, suppose one wanted to refer to a Real number that does not have a "proper name" (unlike, e.g., π); then the class of potentially useful properties is so vast that no REG algorithm can hope to find the right properties. As long as the number of *properties* is *denumerably* infinite, then this problem does not arise, although termination can become problematic: *if* a distinguishing description $[\![P_1]\!] \cap ... \cap [\![P_n]\!]$ exists, *then* the algorithm will find such a description in finite time, because each of the n properties in the Logical Form will be found in finite time; if no distinguishing description exists, however, the algorithm never terminates. To be on the safe side, when we prove completeness, we will assume that the set of properties is at most denumerably infinite, whereas the set of distractors is finite.

Proving Intersective Completeness With these considerations in mind, it becomes possible to prove theorems about intersective completeness for property-based IA. This proves that, unlike IA_{Att}, IA_{Prop} is as powerful as it can be expected to be. (Of course this does not mean that the algorithm must always generate an RE that is natural or useful; see chapter 5 for discussion.)

Theorem 1 Completeness of IA_{Prop}. Suppose there are at most denumerably many properties, and finitely many (one or more) distractors. Then if an object can be individuated by intersecting a finite number of properties, IA will find such an intersection.
Proof: Suppose $[\![Q_1]\!] \cap ... \cap [\![Q_m]\!] = \{r\}$, where properties $Q_1, .., Q_m$ occur in $\mathcal{P}$ in the order indicated by the subscripts. Now either IA returns `Success` *before* it has inspected all of $Q_1, ..., Q_m$, or it reaches the point where all of $Q_1, ..., Q_m$ have been inspected. This does not mean that all of $Q_1, ..., Q_m$ have necessarily been included in L, since other properties in $\mathcal{P}$ may have been selected which cause some of $Q_1, ..., Q_m$ not to remove any distractors. Yet, when all of $Q_1, ..., Q_m$ have been inspected `Success` must have been achieved. To see this, let Des_i be the Logical Form that results after processing (i.e., inspecting and possibly including) Q_i. Then a proof by induction over i shows that $[\![Des_i]\!] \subseteq [\![Q_1]\!] \cap .. \cap [\![Q_i]\!]$, for all $i \leq m$. It follows that $[\![Des_m]\!] \subseteq [\![Q_1]\!] \cap .. \cap [\![Q_m]\!] = \{r\}$. But $r \in [\![Des_m]\!]$, so $[\![Des_m]\!] = \{r\}$. □

Now that we have achieved an understanding of the completeness properties of the Incremental Algorithm, more fine-grained questions appear on the horizon. In particular, we need to discuss how efficient the algorithm is. Efficiency, after all, was one of the motivations behind the algorithm.

4.8 Computational Tractability of REG Algorithms

As was pointed out in the Introduction to this chapter, Dale and Reiter claimed that the IA is superior to its competitors in two respects, namely, that the IA generates more humanlike Logical Forms (see chapter 5), and that it generates these faster. Let us assess the run-time complexity of these algorithms. Let

1. n_a = number of properties known to be true of the intended referent.
2. n_d = number of distractors.
3. n_l = number of attributes mentioned in the RE that is generated.[8]

8 The variable n_l, from [Dale and Reiter, 1995], counts attributes rather than properties (i.e., combinations of an attribute and a value) reflecting Dale and Reiter's tacit assumption that the values of a given attribute cannot overlap (cf., our section 4.7). If this assumption is abandoned, n_l should count the number of properties.

Under Dale and Reiter's analysis, GR has a complexity of $n_a \times n_d \times n_l$, because it needs to make n_l passes through the problem, at each stage checking at most n_a attributes to determine how many of the n_d distractors they rule out. By contrast, if we focus on the problem of finding attributes (leaving the problem of finding an optimal value, and the above-mentioned problems with overlapping values, aside), then IA has a complexity of $n_d \times n_l$, because it requires n_l passes but does not look for the optimal attribute at each stage, since this is fixed in the Preference Order. Dale and Reiter assessed the complexity of FB as $n_a^{n_l}$, so the function grows exponentially. Note that, under a conventional interpretation, this analysis does not so much set aside the IA, but the FB: whereas the latter's complexity is exponential, both GR and IA have polynomial complexity. In other words, there are no strong computational reasons for preferring IA over GR.

A more interesting question is whether computational complexity should be a relevant consideration at all in the evaluation of REG algorithms – or computational cognitive models more generally (cf., section 16.1), for that matter. First, let's view an algorithm as a model of human behaviour. Current REG algorithms seldom pretend to model the *process* (i.e., "How?") aspects of human reference production (section 16.1); at best, they offer a reasonable approximation of the descriptions produced by a human speaker, or of their usefulness. From this perspective, the factors that determine the choice between two algorithms have nothing to do with the speed of these algorithms, but only with the quality of their output. If algorithms aimed to capture the *how*, then REG algorithms could be interpreted as making predictions about speech-onset times, for instance, by tracing how much time each algorithm takes in a given referential situation and comparing this with the time that a speaker takes before starting to produce an RE in that situation.[9] This manner of analysing complexity would be interesting, but this analysis would be starkly different from the type of analysis offered above: from the above perspective, slow means bad; from the suggested perspective, slow might be good (namely, in situations where speakers are slow).

It might be argued that complexity is worth studying for practical reasons, because an intractable algorithms can be too slow to be useful. However, this is debatable as well. Probably no RE contains more than a hundred properties. This realization instantly removes the variable n_l from the complexity of FB,

9 See [Gatt et al., 2012] for a tentative analysis along these lines. See also section 16.3, where some of these arguments are revisited.

causing the algorithm to run in polynomial time. Moreover, a polynomial algorithm whose constants have high values can take more time in actually occurring situations than an exponential one whose constants have low values. It is therefore difficult to assess the practical implications of theoretical complexity results (i.e., ones that use broad complexity classes). Moreover, complexity analyses can omit important aspects of an algorithm: the problem of finding appropriate values (of a given attribute) is often left out of consideration, for instance. Similarly, we shall see in the next chapter that finding a good Preference Order for IA can be very difficult, and this task is not taken into account in the above complexity analysis either.

Computational complexity analyses of REG algorithms have faded into the background in recent years. For example, when REG algorithms were tested in international competitions (as discussed in the next chapter), their run-time behaviour was not assessed. Even though this might change in the future (see section 16.3), this book shall follow the same practice, except where the complexity implications of an algorithm are too stark to neglect.

4.9 Salience

Let us see briefly how REG algorithms can take salience into account. For simplicity, we shall focus on situations in which the speaker and the hearer understand the salience (cf., section 1.6) of each domain object in the same way.

Some accounts of salience treat it as a two-valued, "black-or-white" concept. Algorithms such as the IA can produce reasonable REs "in context" when the set of salient objects is limited in some way, for example, to those entities mentioned in the previous utterance [Passonneau, 1996], [Jordan, 2000c]. But it has long been acknowledged that it is more natural to think of salience — just like height or width, for example — as coming in degrees (e.g., [Alshawi, 1987]). Accordingly, theories of linguistic salience do not merely separate what is salient from what is not; they assign referents to different salience bands, based on factors such as recency of mention and syntactic structure (e.g., [Chiarcos, 2011] for a survey).

One popular approach is to model salience by means of a *focus stack* in the style of Grosz and Sidner [Kronfeld, 1990, Dale and Reiter, 1995, DeVault et al., 2004]: an RE is taken to refer to the highest element on the stack that matches its description. Krahmer and Theune use a more flexible

method, associating a *salience weight* (*sw*) with each object, and interpreting an RE like *the man* as referring to the man with the highest salience weight [Krahmer and Theune, 2002]. They keep the Incremental Algorithm of section 4.6 as it is, except for the start of the algorithm, where they limit the domain to those entities that are salient enough that a generator needs to take notice of them. To do this, they let the algorithm of section 4.6 start with a clause that equates the set of distractors to those that are at least as salient as the referent. The rest of the algorithm stays exactly the same. Given the problems that we encountered with IA_{Att}, we take the property-based IA_{Prop} as the starting point of the algorithm (Algorithm 7).

Algorithm 7 The Incremental Algorithm (based on properties and salience)

Input: A set of domain objects containing a target referent r and a non-empty set M of distractors. The function sw assigns a *salience weight* to each element of the domain. A set $\mathcal{P}$ of properties defined on the domain, where each element of $\mathcal{P}$ holds true of r. Finally, a linear Preference Order defined on $\mathcal{P}$.
Output: A purely conjunctive distinguishing description $\mathcal{D}$ of r if one exists.

1: Start out with an empty $\mathcal{D}$
2: Remove from M all elements d such that $sw(d) < sw(r)$
3: **while** Not all distractors have been ruled out and $\mathcal{P} \neq$ empty **do**
4: Select a new property P from $\mathcal{P}$, *choosing the most preferred one*
5: **if** P is false of some distractors **then**
6: Add P to $\mathcal{D}$
7: Remove P from $\mathcal{P}$
8: Remove from M all distractors ruled out by P

To see how the new clause causes the algorithm to work, suppose the domain consists of 100 mice, with two of them, Minnie and Mickey, more salient than the others. Assume $sw(Mickey) = sw(Minnie) = 10$, but $sw(x) < 10$ for all others, for instance, because the two named mice were encountered more recently than the others. Suppose, furthermore, that Mickey is the target referent, so the relevant (i.e., sufficiently salient) part of the domain is $\{Mickey, Minnie\}$. The task of the algorithm is then to set Mickey apart from Minnie by finding a property that is true of Mickey but not of Minnie (e.g., the property *male*). Other REG algorithms can be modified in the same way, by limiting the domain that the algorithm pays attention to.[10]

10 Alternatively, it is possible to let the set M of distractors remain unchanged from section 4.6, but to modify the algorithm in such a way that insufficiently salient domain elements are

Once differences in salience are taken into account, the question comes up how to choose between different types of NP. When, for example, is it appropriate to use a demonstrative ("this man", "that man") or a pronoun ("he", "she")? As for demonstratives, a body of work by Paul Piwek has shown it to be remarkably difficult to know when these should be used (e.g., [Piwek, 2008]). Regarding pronouns, Krahmer and Theune explored what would happens if "he" was taken to abbreviate "the (most salient) man", and "she" "the (most salient) woman", a move that would bring pronouns into the orbit of REG. Because of their in-built preference for brevity, the resulting algorithms would tend to choose pronouns whenever these can be interpreted unambiguously. (For alternative analyses, see [McCoy and Strube, 1999, Henschel et al., 2000, Callaway and Lester, 2002, Kibble and Power, 2004].)

Most work on REG takes its departure from a Knowledge Base. By contrast, studies focussing on salience frequently use *text* as their starting point [Poesio and Vieira, 1998, Belz et al., 2010, among others], [Siddharthan and Copestake, 2004]. This perspective is also taken by the GREC Evaluation Challenge [Belz et al., 2008] and [Belz et al., 2010], which invited algorithms to choose the type of RE that suits a particular referent in a particular linguistic context. A detailed discussion of the choice between pronouns, demonstratives, and full NPs is beyond the aims of this book (see e.g., section 1.6), but it will be useful to return briefly to the effect of salience on the Content Determination task for full NPs.

Although these approaches make sense, it is worth acknowledging their limitations. It is possible, for instance, to question the claim that only those domain elements need to be considered that are at least as salient as the referent: if the domain contains two men, one of whom is just slightly more salient than the other, then can we really refer to him as "the man"? Perhaps, in other words, the clause $sw(d) \geq sw(r)$ in Algorithm 7 may need to be replaced by one that allows d to have a slightly lower salience weight than r as well.

More importantly, a linguist confronted with Algorithm 7 will point out that salience is not the only factor to be taken into account. Deirdre Wilson offers the following example [Wilson, 1991]:

Sean Penn attacked a photographer. The man was quite badly hurt.

disregarded when the algorithm tests whether all distractors have been ruled out. This alternative version tests whether $M \cap \{x \mid sw(x) \geq sw(r)\} = \emptyset$. The two versions generate the same descriptions in most, but not all, cases; I do not know which one is empirically most accurate.

In this case, Sean Penn is at least as salient as the photographer, yet the latter is more likely to be the intended referent for "the man". In Wilson's view, it is necessary to take into account such factors as (in some cases) the inherent plausibility of the resulting interpretation and (in other cases) the ability of an interpretation to yield inferences that are of interest to the hearer. Both types of explanation, and particular the latter one, with its roots in Relevance Theory (cf., section 3.3), would be extremely difficult to implement in a computer program, because the details of the explanation have not been worked out in sufficient detail. Yet Wilson has a point. At times like this, linguists and computer scientists often part company.

Different standards of precision lie at the heart of many misunderstandings between linguists and computer scientists: frequently the former fail to understand why the latter can accept a relatively simple-minded account (which disregards some important issues), whereas the latter wonder why linguists can be satisfied with an explanation that computer scientists would regard as handwaving (because some crucial things are left undefined). Similar misunderstandings can be found in many other areas of Cognitive Science.

4.10 Summary of the Chapter

We have introduced the classic REG task, which resulted from Dale and Reiter's focus on identification of the referent. We have also introduced some of the best-known algorithms addressing this task, the classic REG algorithms, namely: the Full Brevity (FB) Algorithm, the Greedy Algorithm (GR), and the Incremental Algorithm; other approaches to the classic REG problem will be discussed in chapter 6.

- The first REG algorithm appears to have been designed in the early 1970s by Winograd. In the 1980s, Appelt and Kronfeld (the California School) constructed REG algorithms addressing difficult questions about reference in the context of formal speech act theory. [Section 4.2].
- Around 1990, Dale and Reiter proposed a slimmed-down reference task that is more feasible than the tasks studied by Winograd and Appelt. The new ("classic") task focussed on reference in one shot, to one referent, and by means of non-relational predicates only. Building on Winograd and Appelt's ideas of incremental generation and Rosch-style basic values, Dale and Reiter's Incremental Algorithm (IA) uses a fixed Preference Order to select properties for inclusion into the generated RE. [Section 4.6]

- The Greedy and Incremental Algorithms share a common monotonic structure, whereby properties are added to a Logical Form one by one and never withdrawn. They differ in the manner in which they select the next property to be examined by the algorithm. The monotonic approach to REG will continue to play a role throughout the book. [Section 4.6]
- Unlike Full Brevity and the Greedy Algorithm, the Incremental Algorithm is an intensional algorithm, because it is able to distinguish between properties that are co-extensive. [Section 4.6]
- Under reasonable assumptions, both FB and GR are intersectively complete. The same holds for the property-based version of the Incremental Algorithm, IA_{Prop}. However, it does not hold for the published version of the Incremental Algorithm, which can fail to refer even in situations where it is possible to identify the referent using simple (i.e., logically atomic) properties. [Section 4.7]
- Here, as in other areas of Cognitive Science, analyses of computational tractability have receded into the background in recent years, giving way to empirical studies. [Section 4.8] One type of empirical study, involving what we shall call a semantically and pragmatically transparent corpus, will taken centre stage in the next chapter.
- Salience degrees can be incorporated into standard REG algorithms but it is as yet unclear what is the empirically most accurate way to do this, and linguistic observations in the tradition of Relevance Theory suggests that existing accounts fail to take some crucial factors into account. [Section 4.9] We shall have more to say about salience in section 9.7.
- The generation of relational descriptions, as in the RE "the cup on the table", was first addressed in [Dale and Haddock, 1991] but frequently disregarded in later years. The generation of relational descriptions will be discussed in section 6.4 and discussed in section 6.5 and chapter 10. More implicitly, relational REs will play an important role in chapters 12 and 13.

5 Testing REG Algorithms: The TUNA Experiment

Dale and Reiter's discussion of REG, which was the topic of chapter 4, opened the door to a very focussed study of reference production. By concentrating on a specific version of the reference task, it became possible to test algorithms. This chapter will discuss how their ideas came to be tested, and what one can learn from these tests.[1]

Unlike other parts of the book, this chapter will offer a certain amount of detail concerning experimental setup and statistical analysis: to convey the spirit of empirically informed computational modelling, it seems important to give the reader an impression of the experimental details at least once. Readers who do not care for such things may want to move on to section 5.7, where we draw conclusions from this experiment, or they may skip this chapter entirely.

Around 1990, statistical methods were starting to pervade computational linguistics. These methods were becoming successful in Machine Translation and Speech Recognition, so linguists started to ask themselves methodological questions that had long been dormant, such as, "How do I know that my theory, or my algorithm, is any good?" Such questions only become pertinent once algorithms have reached a certain level of sophistication: to find flaws in a shoddy piece of work, you don't need extensive experimentation; but linguistic theories and algorithms had matured, so evaluation became crucial.

The results of this change are stark. Pick up a conference proceedings in Computational Linguistics around the middle of the 1980s and you'll find clever computing science, frequently linked with Formal Logic and theoretical linguistics; what you will not find is systematic evaluation. Now fast-forward to the year 2000, and Computational Linguistics had changed beyond recognition: conference papers lacking empirical evaluation had become a rarity. Few would argue that this is a bad thing. However, links with logic and linguistics had correspondingly declined [Reiter, 2007]. Computational Linguistics, in other words, had taken a drastic "empirical turn". Our chapter 10 will return to these issues, arguing that logic has a crucial role to play in REG and other areas of Computational Linguistics.

1 I am much indebted to Albert Gatt for providing me with statistical analyses that modify the ones reported in journal articles and conference proceedings, by disregarding reference to sets. The reasons for disregarding sets in the new analyses are explained in sections 5.1 and 5.3, which build on [van Deemter et al., 2012b] and [van Deemter et al., 2012a]. Section 5.9 elaborates on [van Deemter and Gatt, 2007]. The TUNA experiment was proposed in [van Deemter et al., 2006], discussed in [Gatt et al., 2007] and [van der Sluis et al., 2007], and reported more comprehensively in [van Deemter et al., 2012b].

Evaluation is often conducted by means of coordinated public contests in which, following an open call, algorithms are tested against the same data set, using the same evaluation method. It is easy to see the appeal of this idea, which came to be associated with the terms "evaluation campaign" and "evaluation challenge". An evaluation campaign enables a research community to compare the behaviour of a large number of algorithms in detail. Without a campaign this is difficult to do, because researchers often do not have access to the details of other people's algorithms. A collective evaluation campaign allows this to be done in an objective way, orchestrated by a committee representative of the research area as a whole.

Evaluation campaigns involving a shared task had been held in many areas of computational linguistics: The campaigns in Information Extraction (in the MUC Message Understanding Conferences) started as early as 1987; the Information Retrieval challenges (as part of the TREC Text Retrieval Conferences) took off in 1992; in Text Summarization, a coordinated evaluation campaign (associated with the DUC Document Understanding Conferences) was first held in 2001. Around the same time, challenges focussing on Machine Translation (the NIST OpenMT challenges) were starting up. These campaigns may not always have been entirely uncontroversial – there is a risk of focussing too long on one particular kind of data and one particular way of assessing the success of an algorithm – but there is a wide recognition that evaluation campaigns have contributed to the considerable successes achieved in these research areas.

Until 2007, however, no systematic evaluation campaign had ever been held in Natural Language Generation (NLG). Conducting such a campaign is challenging, because NLG systems can take very different types of inputs, which makes it difficult to compare systems on a level playing field. Extensive discussions took place, culminating in a special session at INLG-2006 in Sydney and a workshop on Shared Tasks and Comparative Evaluation in NLG in 2007 [Gatt and Belz, 2010]. REG was singled out as particularly suitable for an evaluation campaign because, thanks to the work of Dale and Reiter, most researchers shared essentially the same assumptions about the enterprise, at least in terms of what the input of the algorithms should be. Yet there were doubters. Some feared that systematic evaluation campaigns would lead to exaggerated competition, and that evaluation campaigns would cause a narrowing of research questions and methods. The advocates of evaluation campaigns won the argument but the doubters ensured that the enterprise was carried out in a spirit of open-minded experimentation, and that subsequent campaigns would use subtly different tasks and a variety of metrics to avoid stasis.

The first NLG Shared Task and Evaluation Challenge took place in 2007, focussing on Attribute Selection for Referring Expressions Generation (ASGRE). They were largely based on the TUNA corpus and evaluation method, on which we focus in this chapter, and known as the TUNA (or ASGRE) challenges. Twenty-two algorithms were submitted to TUNA-REG'07 by 13 research teams [Belz and Gatt, 2007]. The TUNA corpus also featured in some of the tasks organized for the second and third challenge, TUNA-REG'08 and TUNA-REG'09, where the number of submitted systems was even greater [Gatt and Belz, 2010]. A fourth challenge focussed on the generation of referring expressions (REs) in linguistic context [Belz et al., 2008]. Recent NLG evaluation campaigns tend to make reference part of a larger enterprise (e.g., allowing a hearer to find her way in a room). We will focus on reference, motivating and discussing in some detail the experimental study that stood at the cradle of these developments, and whose outcomes, in terms of both annotated corpora and evaluation metrics, lie at the heart of the Challenges.

The plan for this chapter is as follows. Section 5.1 explains the thinking behind the TUNA experiment, after which section 5.2 compares this experiment with its predecessors. Section 5.3 explains how a transparent corpus – a corpus in which text is coupled with the data that it describes – resulted from the experiment, which is analysed in sections 5.4, 5.5, and 5.6. The chapter ends with two different sets of conclusions: one that arises from the original TUNA experiment itself (section 5.7), and one that arises from the evaluation campaigns that it gave rise to (section 5.8).

5.1 Why the TUNA Experiment?

Following earlier studies ([Passonneau, 1995], [Gupta and Stent, 2005]), researchers in the TUNA project compared the IA to its main competitors, Full Brevity and the Greedy Algorithm, using data elicited from human participants in a tightly controlled experiment. The TUNA experiment was set up to test Dale and Reiter's hypothesis, that the IA is superior. Dale and Reiter had cited two reasons. One is that the IA is computationally tractable (cf., section 4.8), and the other is that the REs generated by the IA are better than the ones generated by the other algorithms. Although their paper was a little ambivalent about what this means, it has usually been interpreted in terms of the degree to which the generated REs resemble the ones produced by human speakers. Following [Belz and Gatt, 2008], we shall call this the criterion of

humanlikeness. It is humanlikeness of the Logical Forms generated on which the present chapter focusses. (A different perspective, which emphasizes the usefulness for a hearer, will be pursued in chapters 8 and Part IV.) In the footsteps of Dale and Reiter, we focus on the task of identifying the referent, temporarily disregarding the fact that REs can serve other communicative purposes [Jordan, 2002, Stone et al., 2003].

In one respect, our experiments were less conservative. As we shall see in chapter 8, certain types of references to sets can be generated by variations of the classic REG algorithms. The quality of the resulting REs was tested as part of our general plan, so everything so far published on "the TUNA experiment" has analysed reference to sets as well as reference to singular entities. But even though we did our best to separate the two phenomena, I believe that clarity is served most in this chapter by focussing only on reference to individuals; sets will be discussed in chapter 8. This departure from previous analyses meant that a number of statistical analyses have had to be done afresh.

For generality, we studied two different domain types, involving furniture (the furniture corpus) and photographs of people (the people corpus). The annotated TUNA corpus is available from the Evaluations and Language resources Distribution Agency (ELDA).[2] Our original evaluation focussed on the classic REG algorithms, discussed in chapter 4. In a nutshell:

- Full Brevity (FB) minimizes the number of properties in the Logical Form generated.
- Greedy Algorithm (GR) selects properties one by one, always choosing a property that excludes the largest number of distractors.
- Incremental Algorithm (IA) selects attributes or properties one by one, using a fixed Preference Order.

As noted in section 4.5, Dale and Reiter gave special treatment to the TYPE attribute: if TYPE is not selected by the algorithm, then IA adds the property to the Logical Form at the end of the search process. The same considerations that make this a reasonable move in combination with IA make it a reasonable move in combination with other algorithms as well. We therefore decided to level the playing field by applying the same idea to FB and GR.

2 See also the web page `http://www.abdn.ac.uk/ncs/departments/computing-science/tuna-318.php`.

5.2 How to Test a REG Algorithm?

Psycholinguistic studies of reference production, which have typically sought to test specific hypotheses, are discussed in chapter 3. Here we focus on efforts, usually undertaken by computational linguists, to evaluate an algorithm as a whole. To convey the flavour of this work and the issues that need to be considered, a considerable amount of detail will be necessary.

One of the first evaluations of REG was Passonneau's work on the famous *pear stories* of [Chafe, 1980], which compared a number of classic algorithms with approaches based on Centering Theory [Passonneau, 1995]. Later, Jordan and Walker used the COCONUT corpus (see chapter 6.6), focussing on the role of dialogue history. Similarly, [Gupta and Stent, 2005] carried out an evaluation on the COCONUT and the MAP TASK [Anderson et al., 1991], [Bard, 2007] corpora.[3] The MAP TASK corpus had resulted from an experiment in which an *instruction giver* explains to a *follower* how to follow a route, which is only shown on the map of the instruction giver. The instruction giver tends to use REs to refer to landmarks on the map. Originally, landmarks were labeled with proper names (e.g., "The Pond"), but to facilitate the study of descriptions, a modified corpus called iMAP was created [Guhe and Bard, 2008], in which landmarks were not labelled, so the instruction giver had to invent descriptions [Guhe and Bard, 2012], [Guhe, 2012]. Gupta and Stent used both corpora to compare the Incremental Algorithm to the Greedy Algorithm [Siddharthan and Copestake, 2004]; their evaluation took both Content Determination and Linguistic Realization into account, using a single evaluation metric.

Although these studies offer important insights, they do not directly address our questions. In the map task corpus, for example, most referents are named entities. In the COCONUT corpus, identification was often not the only referential goal of interlocutors, as we have seen. Furthermore, the evaluation metric used by Gupta and Stent incorporated syntactic factors, going beyond the purely semantic task-definition that the IA sought to address. These issues, of course, are only limitations from our present viewpoint.

One study offers a more straightforward comparison of the IA and GR [Viethen and Dale, 2006a]. This study used identification as the sole communicative intention. Algorithms were tested against a small corpus of 118

3 For more about the COCONUT corpus, see section 6.6.

descriptions, obtained by asking experimental participants to refer to drawers in a filing cabinet, which differed on four dimensions, namely, COLOUR, ROW, COLUMN and whether or not a drawer was in a corner. The primary evaluation metric was *recall*, defined as the proportion of descriptions in the corpus that an algorithm reproduced perfectly. The comparison of IA and GR revealed a recall rate of 79.6% for the latter, compared to a 95.1% for the IA. The corpus contained 29 overspecified descriptions, of which the IA was able (given the right Preference Order) to reproduced all but five.

We decided to perform a more extensive evaluation of REG Content Determination in a setting where REs do not rely on linguistic context. A complicating factor is that the IA is a *family* of algorithms, because there are as many versions of it as there are Preference Orders. Considering all of these orders would have meant comparing a huge number of algorithms, with deleterious effects on the reliability of any statistical results. We therefore opted to investigate only orders that are plausible. Where possible, plausibility was defined in terms of earlier psycholinguistic work. The data against which we compared the IA and its predecessors came from the TUNA corpus, which contains descriptions of simple, well-defined objects (artificially constructed images of furniture), and of more complex objects (photographs of people).

Given that Dale and Reiter's claims focussed on Content Determination, our comparison of REG algorithms should disregard differences in lexical choice and Linguistic Realization. Suppose an intended referent has the properties $\langle$type : sofa$\rangle$ and $\langle$colour : red$\rangle$, and two people produce the descriptions "the settee which is red", and "the red sofa", respectively. An algorithm that selects each of the two properties above should be counted as achieving a perfect match with both descriptions. In line with Dale and Reiter's starting point, the corpus-based evaluation on which we report here focusses on an assessment of the *humanlikeness* of the Logical Forms generated by a given REG algorithm. In other words, we ask how well an algorithm mimics speakers.

How to measure the success of a generated RE? A measure such as recall does not take into account the degree of overlap between two descriptions. Consider a case where a human-produced description expresses the attributes $\{$COLOUR, SIZE$\}$, whereas an algorithm outputs $\{$COLOUR$\}$ only. Recall would simply count this as a mismatch, ignoring the overlap between the two attribute sets. Instead, we adopted the Dice coefficient, a well-accepted metric that computes the degree of similarity between two sets in a straightforward way. Dice is computed by scaling the number of attributes that two

descriptions have in common, by the overall size of the two sets:

$$Dice(D_H, D_A) = \frac{2 \times |D_H \cap D_A|}{|D_H| + |D_A|} \tag{5.1}$$

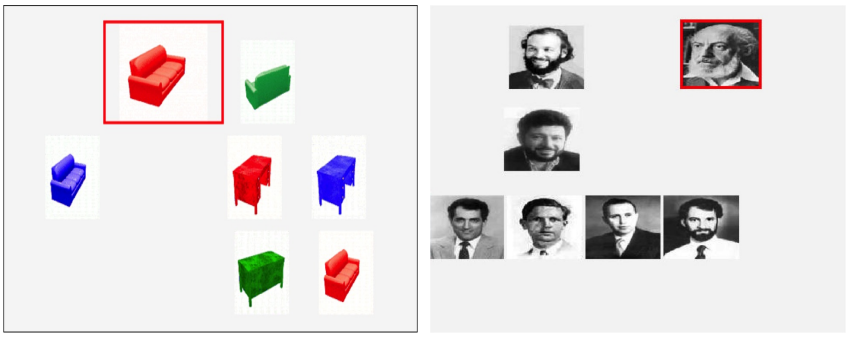
D_H is the set of attributes expressed in the description produced by a human author and D_A is the set of attributes expressed in the Logical Form generated by an algorithm. Dice yields a value between 0 (no agreement) and 1 (perfect agreement). As an indicator of the overall recall of an algorithm, I will also report the *perfect recall percentage* (PRP), the proportion of times the algorithm achieves a score of 1, agreeing perfectly with a human author on the semantic content of a description. After all, one might take the view that to get it slightly wrong is to get it wrong.

5.3 The TUNA Corpus and Its Annotation

The experiment was carried out over the internet over a period of three months, yielding 2280 singular and plural descriptions by 60 participants. Later, two smaller data sets have been constructed using the same overall methodology for TUNA-REG'08. Algorithms were compared against human descriptions in the original TUNA corpus. These results were subsequently validated against one of the TUNA-REG'08 test data sets; because this validation did not reveal any new insights, however, I will not discuss it in detail here, referring the reader to [van Deemter et al., 2012b] for details. Another TUNA-REG'08 test set, however, was used as development data, in a way that will become clear soon. We let participants refer to objects in two domain types, yielding the *furniture corpus* and the *people corpus*.

We had to find a setting in which large numbers of human descriptions could be obtained, each referring to an object for the first time. We chose an experiment in which each trial consisted of one or two target referents and six distractor objects, with the targets clearly demarcated by red borders, as illustrated in Figure 5.1. Objects were displayed in a sparsely populated 3 (row) by 5 (column) matrix; their positioning inside this invisible matrix was determined randomly at runtime, for each participant and each trial.

Participants were asked to identify the target in each trial. They were told they would be interacting with a language-understanding program that would interpret their description and remove the referents from the domain. The referent was automatically removed from the domain after a participant had entered a description. For added realism, the system removed the correct referent(s) on



(a) Furniture trial (b) People trial

Figure 5.1
Trials in the TUNA *elicitation experiment.*

Attribute	Possible Values
TYPE	*chair, sofa, desk, fan*
COLOUR	*blue, red, green, grey*
ORIENTATION	*front, back, left, right*
SIZE	*large, small*
X-DIMENSION (column number)	$1, 2, 3, 4, 5$
Y-DIMENSION (row number)	$1, 2, 3$

Table 5.1
Attributes and values in the furniture *corpus*

75% of the trials, but the wrong one(s) on a (randomly chosen) quarter of trials. With hindsight, this setup may have introduced confounding factors in our design. For example, participants' referential behaviour may have altered after the system "misinterpreted" a description, causing less risk-taking behaviour [Carletta and Mellish, 1996b]. It was for these and other reasons that we used the TUNA-REG'08 test data as a second, validating test set, in which this problem did not arise.

The furniture corpus consists of references in domains constructed using pictures of furniture and household items obtained from the Object Databank, a set of stylize, digitally created images developed by Michael Tarr and colleagues at Brown University. For each object, six pictures are provided, representing the same object at six different orientation angles. Four types of objects were selected from the Databank. For each object, there were four versions

corresponding to four different values of ORIENTATION. Pictures were manipulated to create a version of each TYPE $\times$ ORIENTATION combination in four different values of COLOUR and two values of size, as shown in Table 5.1. As shown in the table, there are two additional attributes, X-DIMENSION and Y-DIMENSION, which describe the location of an entity in the 3×5 grid.

The people corpus consists of references elicited in domains consisting of high-contrast, black-and-white photographs of people, which had been used in [van der Sluis and Krahmer, 2004]. This corpus is more complex than the

Attribute	**Possible Values**
TYPE	*person*
ORIENTATION	*front, left, right*
AGE	*young, old*
BEARD	0 (`false`), 1 (`true`), *dark, light, other*
HAIR	0 (`false`), 1 (`true`), *dark, light, other*
HASGLASSES	0 (`false`), 1 (`true`)
HASSHIRT	0, 1
HASTIE	0, 1
HASSUIT	0, 1
X-DIMENSION (column number)	$1, 2, 3, 4, 5$
Y-DIMENSION (row number)	$1, 2, 3$

Table 5.2

Attributes and values used in the people corpus

furniture corpus, because a given portrait can be described using many different attributes (e.g., "the bald man with the funny smile and the nerdy shirt"). In Van der Sluis and Krahmer's study, three attributes – BEARD, HASGLASSES and AGE – had been particularly frequent, therefore we constructed each people domain in such a way that the target(s) could be distinguished uniquely from their distractors using a combination of these three.

The experiment consisted of 38 experimental trials, divided into 20 furniture trials and 18 people trials, each with one or two targets and six distractors in the sparse matrix. Each trial displays what we call a *domain*. For furniture, the domains were constructed by taking each possible combination of attribute-value pairs in each domain type, and constructing a domain in which that combination was a minimally distinguishing description (briefly: the minimal description) of the referent(s), by which we mean a distinguishing description containing a set of n properties, where no distinguishing description of the referent exists that contains fewer than n properties. The available attribute-value

pairs in a domain type were represented an approximately equal number of times. For example, of the 12 furniture domains in which ORIENTATION was part of the minimal description, a target faced *front* or *back* exactly half the time, and *left* or *right* in the rest.

Table 5.3 summarizes the experimental design, which manipulated one within-subjects and two between-groups factors. The within-subjects factor manipulated the Cardinality and Similarity of objects. In the case of plural domains with two target referents, the two referents may or may not be sufficiently similar to be describable by means of the same minimal conjunction of properties. Where this is not possible, one may have to split the description in two parts, as in *the red table* and *the blue sofa*, for example. Accordingly, Cardinality/Similarity had three levels:

1. **Singular** (SG): 7 furniture domains and 6 people domains contained a single target.
2. **Plural/Similar** (PS): 6 furniture domains and 6 people domains had two referents with identical values for the attributes with which they could be distinguished from their distractors. For example, two pieces of furniture might both be blue in a domain where all distractors are red. In the furniture domain type, the two referents in this condition had different values of TYPE (e.g., one was a chair, and the other a sofa), but in the people domain type, they were identical (because all entities were men).[4]
3. **Plural/Dissimilar** (PD): In the remaining 7 plural furniture trials and the 6 plural people trials, the targets had different values for the minimally distinguishing attributes. Thus, plural descriptions in this condition would always involve disjunction if they were to be distinguishing.

	Furniture			**People**		
	SG	PS	PD	SG	PS	PD
-LOC (N = 30)	525	450	525	450	525	525
+LOC (N = 30)	210	180	210	180	210	210

Table 5.3

Experimental design and number of descriptions within each cell. SG *stands for singular,* PS *for Plural/Similar, and* PD *for Plural/Different.*

Half of the participants were discouraged from using locative expressions (−LOC condition), whereas the other half (+LOC) were not. The former were told that the language understanding program they were interacting with had different information about the position of objects, so using locatives would

4 This design made Plural/Similar references to furniture more complex than to people. This is another reason why the present analysis focusses on reference to singular entities, where this problem does not exist.

be counterproductive. Participants in +LOC were told that the system had access to the complete domain of objects, including location. However, locatives were not included in Dale and Reiter's discussion of REG algorithms. Moreover, location requires extensions to the original IA to deal with relations such as *above* [Gorniak and Roy, 2004, Kelleher and Kruijff, 2006b]. For these reasons, locative descriptions will not be discussed here.

Participants were randomly assigned to a condition and read the corresponding instructions. They were asked to complete the experiment (i.e., all 38 furniture and people trials) in one sitting. The instructions emphasized that the purpose of their descriptions must be to identify referents. Trials were presented in randomized order. Each trial consisted of a presentation of a domain, as shown in Figure 5.1, where participants were prompted for a description of the target referent(s). This was followed by a feedback phase, in which the system removed the target referent. A total of 60 participants completed the experiment, 15 in each group depicted in Table 5.3.

Figure 5.2 shows how our corpora were annotated. Each corpus description was paired with the domain in which it was produced. Each description was represented in three different ways: (a) the original string typed by a participant (the `STRING-DESCRIPTION` node); (b) the same string with all substrings corresponding to an attribute, annotated using `ATTRIBUTE` tags (the `DESCRIPTION` node); and (c) a simplified representation consisting only of the set of attributes used by a participant (the `ATTRIBUTE-SET` node). Evaluation required that the domains on which human and algorithm had produced references be compatible whenever possible. The exception was when expressions contained attributes that were not specified in the domain at all (e.g., where a person was described as being *serious*), something that is unavoidable in complex domains of this kind; these were tagged using `name='other'`. In the evaluations reported in Sections 5.4 and 5.5, these attributes were discounted when algorithms were compared.

Two of the authors of the study annotated about 50% of the data, which was then cross-checked by the other. Disagreements were discussed until a consensus was reached. The annotated text was processed automatically to produce the XML representation discussed above. The reliability of our annotation scheme was evaluated by comparing a subset of 516 descriptions in the corpus to the annotations made by two independent annotators who used the same annotation manual. Our findings confirmed that the three sets of independently annotated descriptions agreed substantially with each other [van Deemter et al., 2012b].

```
<TRIAL ID='s2t3' (...)>
    <DOMAIN>
       <ENTITY type='target'>
          <ATTRIBUTE name='type' value='sofa' />
          <ATTRIBUTE name='colour' value='red' />
          <ATTRIBUTE name='orientation'  value='right' />
          <ATTRIBUTE name='size' value='large' />
          <ATTRIBUTE name='x-dimension' value='1' />
       </ENTITY>
       <ENTITY type='distractor'>
          <ATTRIBUTE name='type' value='sofa' />
          <ATTRIBUTE name='colour' value='red' />
          <ATTRIBUTE name='orientation' value='left' />
          ...
       </ENTITY>
       ...
    </DOMAIN>

    <STRING-DESCRIPTION>
       the sofa facing right
    </STRING-DESCRIPTION>

    <DESCRIPTION>
       the
       <ATTRIBUTE ID='a1' name='type' value='sofa'>
          sofa
       </ATTRIBUTE>
       <ATTRIBUTE ID='a2' name='orientation' value='right'>
          facing right
       </ATTRIBUTE>
    </DESCRIPTION>

    <ATTRIBUTE-SET>
       <ATTRIBUTE ID='a1' name='type' value='sofa'/>
       <ATTRIBUTE ID='a2' name='orientation' value='right'/>
    </ATTRIBUTE-SET>
</TRIAL>
```

Figure 5.2

A corpus instance ("the sofa facing right"). Adapted from [van Deemter et al., 2012b].

5.4 Analysis of the Furniture Corpus

Testing all Preference Orders would not be practical. For this reason, we used our development data from TUNA-REG'08 to estimate the probability with which different attributes are produced and we used these probabilities to determine a set of Preference Orders.

We start with the furniture corpus. As we have seen, the psycholinguistic literature suggests that, of the three attributes in this domain, COLOUR will tend to be strongly preferred, whereas SIZE is dispreferred [Pechmann, 1989, Belke and Meyer, 2002]. The situation is far less clear with ORIENTATION.

Attribute	Frequency (%)
TYPE	56 (31.6)
COLOUR	49 (27.7)
ORIENTATION	20 (11.3)
SIZE	20 (11.3)
Y-DIMENSION	16 (9)
X-DIMENSION	13 (7)
OTHER	3 (1.7)

Table 5.4

Frequency of attribute usage in the TUNA-REG'08 *development data for the furniture corpus. (Locative attributes and* OTHER *are ignored in the present study.) Frequency says how often the different attributes occurred in the corpus as a whole, for instance,* 31.6% *of attribute occurrences in the corpus were types.*

Frequencies computed from our development data set, displayed in Table 5.4, confirm the predicted trends, while showing a tie between SIZE and ORIENTATION. We therefore expected that Preference Orders that put COLOUR first will generally perform better (e.g., have higher Dice scores). We can test these hypotheses by comparing all 6 possible IAs. The one with the order COLOUR $>$ ORIENTATION $>$ SIZE (which we call IA-COS), the one with the order COLOUR $>$ SIZE $>$ ORIENTATION (which we call IA-CSO), and so on.

Table 5.5 shows the performance of each algorithm; for completeness it shows not only results computed on the original TUNA data set, but results computed on the TUNA-REG'08 data as well; the mean Dice scores on the two data sets are strongly correlated. (Comparing only the means obtained on the singular descriptions, $r_9 = .985; p < .001$.) The ranking of the algorithms is

	Original TUNA Data		TUNA-REG'08 Data	
	Mean (SD)	PRP	Mean (SD)	PRP
IA-COS	0.917 (.12)	60.9	0.916 (.16)	69.1
IA-CSO	0.917 (.12)	60.9	0.916 (.16)	69.1
RAND	0.840 (.15)	31.4	0.826 (.18)	34.6
IA-OCS	0.829 (.14)	25	0.829 (.15)	25.5
IA-SCO	0.815 (.14)	19.2	0.823 (.15)	18.2
IA-OSC	0.803 (.16)	22.4	0.801 (.17)	25.5
IA-SOC	0.780 (.16)	18.6	0.782 (.16)	18.2
FB	0.841 (.17)	39.1	0.845 (.17)	37.5
GR	0.829 (.17)	37.2	0.845 (.17)	37.5

Table 5.5

Mean Dice scores and standard deviations for the furniture corpus, in the original TUNA *Data and the* TUNA-REG'08 *Data set that was used for validation, with* PRP *scores per algorithm. All figures are for singulars only. The letters* C, O, *and* S *stand for colour, orientation, and size.* PRP *stands for* perfect recall percentage. *As before,* FB *is the Full Brevity algorithm,* GR *the Greedy Algorithm, and* IA *the Incremental Algorithm (with different Preference Orders).*

largely the same for the two data sets, particularly for the top and bottom rankings. The overall rankings of the algorithms on the TUNA data also correlated significantly with the rankings on the TUNA-REG'08 data (focussing on the singular descriptions $\rho_9 = .98; p < .001$). This suggests that the two data sets are largely compatible.

Note that the two top IAs are the ones that place COLOUR first. A univariate ANOVA test was conducted to compare all the versions of the IA on the original TUNA data. This analysis showed up a highly significant main effect of ALGORITHM: $F(6, 1092) = 22.697, p < .001$), so meaningful comparisons between the different algorithms can be made. A standard way to do this is to use Tukey's Honestly Significant Differences.

Pairwise comparisons using this method yielded the four homogeneous subsets of algorithms (A,B,C,D) displayed in Table 5.6. Algorithms in the same subset (i.e., which share a letter) are statistically indistinguishable from each other at the level $\alpha = .05$, in other words, the differences between them that we measured were small enough that they may have been accidental. (For example, the difference between the Dice score of IA-RAND, at 0.840, and the Dice score of IA-OCS, at 0.829 could be accidental.) Algorithms that do not share a letter had scores that were significantly different from each other. The table

IA-COS	A			
IA-CSO	A			
IA-RAND		B		
IA-OCS		B		
IA-SCO		B	C	
IA-OSC			C	D
IA-SOC				D

Table 5.6

Homogeneous subsets among versions of the IA *in the furniture corpus. Algorithms that do not share a letter are significantly different at* $\alpha = .05$.

shows a separation between the two IAs that prioritize COLOUR, and the other algorithms. Clearly, even in a small domain with few dimensions of variation among objects, the human-likeness of the output of the IA is affected by the Preference Order. For the next part of our analysis, we will compare GR and FB against an optimal IA, namely, IA-COS, and the non-optimal IA-SOC.

Table 5.5 shows that the brevity-oriented Greedy and Full Brevity algorithms fall somewhere between the two top-scoring IAs that place COLOUR first, and the others, with IA-RAND actually outperforming them by a small margin in terms of mean scores. We conducted a separate univariate ANOVA comparing the best and worst IAs to FB and GR, which showed up a significant main effect of ALGORITHM ($F(3, 624) = 20.559, p < .001$).

IA-COS	A		
FB		B	
GR		B	
IA-SOC			C

Table 5.7

Homogeneous subsets among the best and worst IA*s with* FB *and* GR. *Algorithms that do not share a letter are significantly different at* $\alpha = .05$.

The post-hoc Tukey's test (Table 5.7) shows that the two brevity-oriented algorithms scored reliably worse than the best IA but reliably better than the worst IA. In other words, Dale and Reiter's prediction, to the effect that an incremental strategy would improve humanlikeness, was only valid for a specific subset of all the possible IAs. These results suggest that Intrinsic Preference is an important component of any analysis that seeks to test Dale

and Reiter's claims. Good Preference Orders outperform all other algorithms, whereas an order that reverses known human attribute preferences does far worse than any other algorithm. Such statements also depend on the evaluation metric used. With the exception of IA-SOC, the overall means in Table 5.5 range between 0.75 and 0.9. These differences seem intuitively small, which might be explained by the simplicity of the domains.

5.5 Analysis of the People Corpus

To see what happens in a more challenging referential domain, we now turn to the people corpus. Given that the number of attributes is greater than in our previous analysis, it is even more crucial to have an *a priori* estimate of what Preference Orders might constitute good IAs. This time, however, the psycholinguistic literature provides little guidance: not a lot has been published on the perceptual salience of attributes such as HASGLASSES. Therefore, we have to rely on frequencies based on the development data set (Table 5.8).

Attribute	Frequency (%)
TYPE	55 (29.6)
HASBEARD (B)	36 (19.4)
HASGLASSES (G)	25 (13.4)
HASHAIR (H)	22 (11.8)
AGE (A)	14 (7.5)
HASSHIRT (S)	4 (2.2)
HASSUIT (S)	3 (1.6)
HASTIE (T)	2 (1.1)
ORIENTATION (O)	1 (0.5)

Table 5.8

Frequency of attribute usage in the development data for the people corpus. (Locative attributes and OTHER *are ignored in the present study.) Frequency says how often the different attributes occurred in the corpus as a whole, for instance,* 29.6% *of attribute occurrences were types.*

The table suggests a gap between the three attributes BEARD, HASGLASSES, and HAIR, and all the others.

To construct different versions of the IA, we took all possible permutations of these three attributes, imposing a fixed order on the other six. Additionally,

we again used a version of the IA that randomizes the Preference Order (IA-RAND) and one that reversed the hypothesized best orders; this is our predicted bad version, IA-SSTAOHGB:

IA-GBHOATSS: HASGLASSES > BEARD > HAIR > ... > HASSUIT

IA-GHBOATSS: HASGLASSES > HAIR > BEARD > ... > HASSUIT

IA-BGHOATSS: BEARD > HASGLASSES > HAIR > ... > HASSUIT

IA-BHGOATSS: BEARD > HAIR > HASGLASSES > ... > HASSUIT

IA-HBGOATSS: HAIR > BEARD > HASGLASSES > ... > HASSUIT

IA-HGBOATSS: HAIR > HASGLASSES > BEARD > ... > HASSUIT

IA-SSTAOHGB: HASSUIT > HASSHIRT > HASTIE > AGE > ... > BEARD

	Original TUNA Data		TUNA-REG'08 Data	
	Mean (SD)	PRP	Mean (SD)	PRP
IA-GBHOATSS	0.844 (.17)	44.7	0.811 (.17)	33.9
IA-BGHOATSS	0.822 (.17)	36.4	0.797 (.17)	32.1
IA-GHBOATSS	0.776 (.21)	29.5	0.77 (.18)	26.8
IA-BHGOATSS	0.728 (.19)	15.9	0.792 (.17)	30.3
IA-HGBOATSS	0.688 (.18)	3.8	0.765 (.17)	25
IA-HBGOATSS	0.658 (.20)	4.5	0.752 (.17)	23.2
IA-RAND	0.598 (.23)	11.4	0.527 (.21)	0
IA-SSTAOHBG	0.344 (.11)	0	0.344 (.08)	0
FB	0.764 (.23)	34.1	0.642 (.23)	19.6
GR	0.693 (.20)	8.3	0.642 (.23)	19.6

Table 5.9

Mean Dice scores and standard deviations for the people corpus, with PRP *scores per algorithm. All figures concern singulars only.*

Table 5.9 shows the overall Dice scores to be more broadly distributed than before, with IA-RAND and IA-SSTAOHBG scoring at or below .6. The worst IA has an extremely low PRP, scoring 0 in the singular data (so it does not match with any of the descriptions). A univariate ANOVA again showed a highly significant main effect of ALGORITHM: $F(7, 1056) = 96.691, p < .001$).

Table 5.10 shows that the two best-performing algorithms differ significantly from all other algorithms. At the bottom of the table, two distinct subsets identify the worst-performing algorithms, one of which is IA-RAND. We conclude that, in the people corpus, there is a strong dependency of the IA on the Preference Order. Once again, the versions of the IA that perform best are those that prioritize frequent attributes.

IA-GBHOATSS	A					
IA-BGHOATSS	A					
IA-GHBOATSS		B				
IA-BHGOATSS			C			
IA-HGBOATSS			C	D		
IA-HBGOATSS				D		
IA-RAND					E	
IA-SSTAOHBG						F

Table 5.10

Homogeneous subsets among versions of the IA *in the people corpus. Algorithms that do not share a letter are significantly different at* $\alpha = .05$.

As before, the means for GR and FB in Table 5.9 fall somewhere between the best- and worst-performing IAs, with FB outperforming GR. We compared these two algorithms to one of the best IAs (IA-BGHOATSS) and the worst (IA-SSTAOHBG). A univariate ANOVA showed a main effect of ALGORITHM ($F(3, 528) = 187.570, p < .001$). The results confirm that FB outperformed GR, and that both are significantly better than the worst version of the IA.

IA-BGHOATSS	A			
FB		B		
GR			C	
IA-SSTAOHBG				D

Table 5.11

Homogeneous subsets among the best and worst IA*s with* FB *and* GR *in the people corpus. Algorithms that do not share a letter are significantly different at* $\alpha = .05$.

5.6 Modelling a Plurality of Speakers

We have so far judged an algorithm by its *average* Dice score, calculated over all the expressions produced by all participants in the TUNA experiment. This strategy masks important issues. It might be, for example, that an algorithm with a poor Dice score offers a near-perfect model of some subjects (but a terrible model of most of them), in which case such an algorithm could be

argued to be better than its average score suggests. It might be well placed to pass the Turing test, for instance (which is based on people's ability to tell whether a certain behaviour was produced by a real person or by a computer). Here we address the issue by charting *post hoc* to what extent the answer to our research question (i.e., how good a model of human referential behaviour the IA is) depends on the speaker on which one happens to focus.

One reasonable question is: Did our participants differ in terms of which algorithm matches them best? We write "s selects algorithm A" as short for "algorithm A has an average Dice match to author s's descriptions that is at least as good as that of all other algorithms considered". If several algorithms have the same average match with a participant's descriptions, then we will say that all of them are selected by that participant. Thus, our question is: *Do different participants select different algorithms?*

In the furniture corpus, variation between speakers were limited: 17 out of 19 select IA-COS and IA-CSO as their best algorithms. The remaining 2 select IA-COS and IA-CSO as their (equal) second choice. More substantial differences were found in the people domain (Table 5.12): there is a clear majority for IA-GBHOATSS, but as many as 7 of the authors select 4 other algorithms, namely, FB, IA-GHBOATSS, IA-BGHOATSS, and IA-HGBOATSS. Interestingly, FB matches a few authors better than any IA. This is only a post-hoc analysis,

	Best	**2nd Best**	**Worst**
FB	2		
IA-GBHOATSS	14	5	
IA-GHBOATSS	1	2	
IA-BGHOATSS	2	10	
IA-HGBOATSS	1		
IA-SSTAOHBG			20
IA-BGHOATSS & IA-BHGOATSS & IA-HGBOATSS	1		
RAND & GR & IA-BGHOATSS		1	
GHBOATSS & IA-BGHOATSS		1	

Table 5.12

Numbers of subjects that had a particular algorithm as their best, second-best and worst match in the **people** *domain (singulars only). Algorithms that were neither the best match nor the second-best match nor the worst match to any speaker are omitted.*

yet it seems striking that even a simple reference task gave rise to such marked differences between speakers. Variation between speakers will be discussed further in section 6.3.

5.7 Lessons from the TUNA Experiment

The research community has been able to learn some useful lessons from the TUNA experiment, some of which are starting to be reproduced for languages other than English (for Dutch in [Koolen et al., 2009]; for Arabic in [Khan, 2015]). Let us start by summarizing them.

The aim of the TUNA experiment was to test the hypothesis that the IA produces Logical Forms that are more "humanlike" than its main competitors. The IA has undeniable strengths, but in the TUNA corpus, it proved to be difficult to confirm this hypothesis unambiguously. For although there always existed a version of the IA that outperformed all other algorithms examined, its success depended substantially on what Preference Order is chosen: a suboptimal Preference Order results in Logical Forms that are significantly worse than those produced by FB and GR, for example. Some human speakers, in fact, will be modelled more accurately by FB and GR than via any incremental generation strategy: in the people data, FB agreed perfectly with a human author about 61% of the time. A practitioner in Language Technology who is looking for a REG algorithm for an unfamiliar application domain might be forgiven for choosing FB or GR, instead of an unproven version of the IA.

Although psycholinguistic findings can sometimes help to find good Preference Orders, in many domain types such principles are of limited help, as we have seen. However, in the absence of psycholinguistic information, the frequency of each attribute in a corpus can help. The TUNA corpus, and the replication of it that we used to predict Preference Order (Section 5.3), give us a relatively strong starting point, because these corpora result from a *semantically balanced* experiment, which guarantees that the most important types of situations occur equally frequently. Most generally available corpora (such as the BNC) are often not balanced in this way. An attribute may occur frequently, in such a corpus, because it happened to have a high discriminatory value in situations that occur often in the corpus domain. A frequency-based strategy could cause this attribute to be over-used in situations where its discriminatory value is lower.

We have seen that the shortcomings of the IA became particularly noticeable in connection with the more complex of the two domain types (i.e., the people domain). This should give pause for thought. In many ways, the people domain was still comparatively simple. People in the real world, as opposed to mere photographs, would have been identifiable in terms of their physical features,

past actions, ideas, and so on. It is unclear whether any of the algorithms discussed here would do well in such situations. As we shall see in Part IV of the book, there are domain types in which the IA performs poorly regardless of Preference Order. These results suggest that future research in REG should focus on the complexities posed by the large and complex domains that speakers are often faced with in real life.

A Letter to the Editor of *Cognitive Science* that discussed our results wondered how much data is needed to find a "good" Preference Order [Krahmer et al., 2012]. Analogous to our section 5.5, the authors sought to determine Preference Orders by counting the frequencies of attributes in the TUNA corpus. They found that, for furniture, tiny samples suffice to construct an IA that performs as well as their best-performing IA; for the people corpus, the results were more varied, although small samples still tended to perform well. Their procedure went like this: they computed Dice and PRP scores for samples consisting of 1, 5, 10, 20, 30, 40, 50, and 150 descriptions, that is, at 8 different levels. Larger samples might be expected to lead to better IAs of course, because they give a more accurate picture of language use. The authors show that for the furniture corpus, a "ceiling" is reached with as few as 5 descriptions; for the people corpus, a Dice ceiling is reached at 10 descriptions, whereas a PRP ceiling is reached at 40. The ceiling is defined as the lowest level that did not score significantly worse than the IA associated with the highest level (i.e., level 8). The fact that, at least in the TUNA domains, a few examples suffice to find a "good" Preference Order does not tell us how much data would be needed to allow IA to beat its competitors [van Deemter et al., 2012a]. The findings of Krahmer and colleagues are useful nonetheless, because transparent corpora such as the full TUNA corpus are very labour-intensive to design.

5.8 Lessons from the TUNA Evaluation Challenges

We have seen that the TUNA experiment is closely connected to three concerted evaluation campaigns. What can one learn from them?

On balance, the NLG evaluation challenges have been a force for good. The community has not been torn apart by strife, and a perusal of the proceedings of recent NLG conferences does not suggest an obvious narrowing of research issues and research methods. Moreover, thanks perhaps to the doubters' warnings, no single evaluation metric has been treated as sacrosanct. It is reassuring to see what variety of metrics was used for the last GIVE challenge,

which focussed on Direction Giving [Koller et al., 2010b]. The metrics ranged from objective criteria (task completion rate, speed, distance travelled, instructions per second, etc.) to subjective criteria based on questionnaires, in which respondents were asked to indicate their agreement with statements such as "Overall the system gave me good instructions", "I was confused about which direction to go in", "I could easily identify the buttons the system described to me, "The system was very friendly", and so on [Koller et al., 2011]. More NLG evaluation challenges are being organized, addressing various NLG tasks. At least for the moment, evaluation challenges are in vogue.

With hindsight, many aspects of the TUNA challenges have been vindicated. In particular, statistical analysis of the results offers additional support for the use of a plurality of quality metrics. As Belz and Gatt wrote, "our results raise the possibility that automatic, corpus-based metrics of human-likeness focus on very different aspects of the quality of human REs than those tapped into by extrinsic, task-based measures." [Gatt and Belz, 2010]). Yet, it seems to me that TUNA's reference task was too easy in terms of the number of distractors and the number of properties that need to be taken into account, and this may have caused researchers to over-rely on tried-and-tested methods. Reference in real life is often far more complex than in the TUNA experiment: one wonders how difficult it is to find "good" Preference Orders when we point out a person in a crowd, or when two people talk about a place they once visited together, or when a researcher talks about a linguistic phenomenon. Reference in such cases is immeasurably more complex than the task of picking out one person among just 6 distractors. To chart this complexity, Part IV of this book will concentrate on situations that are a bit closer to real life.

The corpora that were used in the challenges should not be taken as the ultimate yardstick by which any REG algorithm should be judged. First, when a corpus has been around for a while, algorithms may be designed with these corpora in mind, hence good performance on these same corpora is no guarantee for good performance in general. Second, these corpora may not contain the data necessary for putting your algorithm to the test. Suppose, for example, your algorithm was designed with the primary aim of producing good descriptions of domain objects that cannot be identified uniquely. Under these circumstances, good performance on TUNA's furniture corpus does not tell you anything useful. To find out whether your algorithm has achieved its aim, you need to conduct a new experiment. Indeed, you may want to test specific hypotheses rather than compute average Dice scores; your work will come to resemble a classic exercise in psycholinguistics (chapter 3).

The evaluations discussed in this chapter have focussed on the extent to which the Logical Forms produced by an algorithm match the ones that employed by human speakers. This is an interesting perspective, which might help computer programs to pass the Turing test [Turing, 1950]; moreover, there is some evidence that, by making REs resemble the ones produced by human speakers, the resulting expressions are more easily understood [Campana et al., 2004]. However, there should also be room for alternative, utility-driven evaluation methods. After all, (cf., section 1.5) human-likeness is not necessarily the most practically useful criterion for judging a production algorithm. Later chapters will often explore a different perspective, relating to the utility of the descriptions generated.

5.9 A Note on Alternative Metrics

The Dice metric was a natural choice for measuring similarity between sets, yet alternative metrics might suit future REG evaluations even better. One problem is that Dice regards all properties as equidistant. Suppose a person referred to an object as "the recently-published book". Now suppose algorithm A referred to it as "the recently-bought book", whereas algorithm B referred to it as "the book on the top shelf". Dice would assess both algorithms as equally faithful to the person's utterance, although intuitively, A beats B, because the properties it selects are more closely related to the human choice. Future research could seek to formalize this idea, perhaps by measuring the similarity between *properties* in terms the similarity of their extensions, for which the Dice metric might then be used once again.

Dice looks at properties individually, without taking their joint effect into account. This is like regarding the choice of a car as if it consisted of a number of independent choices: for the wheels, for the chassis, and so on. Like car parts, the properties in an RE should fit together. This can sometimes be a matter of style, as when two properties represent different views of the referent (section 8.7). In other cases, "fitting together" can have a more easily measurable meaning. Suppose algorithms C and D refer to a target using different descriptions, each deviating from a human-produced description by only one property. But whereas C produces a distinguishing description, D produces a description that fails to identify the referent. Dice would assign equal scores to C and D yet, functionally, C resembles the human-produced description more closely. Finding better metrics seems an important area for further research.

5.10 Summary of the Chapter

We have discussed an empirical investigation of the classic REG algorithms, asking how similar the output of these algorithms is to the REs produced by people. The TUNA experiment was discussed in detail, and so was its role in a series of evaluation campaigns. As for the TUNA experiment itself,

- The TUNA experiment dealt with the classic REG problem, focussing on one-shot, one-referent REs with just one argument place. [Section 5.3]. Moreover, it focussed on small domains involving just one or two referents and 6 distractors, and this may have limited its "ecological validity" (section 3.7).
- Existing REG algorithms are good at mimicking human behaviour in simple situations. [Section 5.4] However, comparisons between the furniture and the people domain suggest that REG algorithms perform worse on the latter. [Section 5.7]. This suggests that these algorithms might struggle in complex referential situations. Parts III and IV of this book will turn to these.

In section 5.8 we have seen that, as for the TUNA evaluation campaigns,

- The TUNA evaluation challenges have helped create common ground between research groups worldwide, although it is less clear that they have boosted the development of new REG algoritms.
- Follow-ups to these evaluation challenges have started to investigate the role of reference as part of a wider communicative task, such as Direction Giving (in the GIVE task, cf., chapter 12).
- Recent evaluation challenges in Natural Language Generation show an awareness that no single metric should be seen as the right one, necessitating the open-minded use of a range of different metrics.

This chapter has shown in some detail how data can be employed to test an algorithm. Data can also be employed to *inspire* an algorithm, for example by means of Machine; we shall soon see that the TUNA corpus has been employed in this way as well.

6 Probabilistic and Other Alternatives to the Classic REG Algorithms

The previous chapters have focussed on the mainstream of work on the generation of referring expressions, where a group of "classic" REG algorithms – which operate by monotonically adding properties until the referent has been identified – took up a central position. Yet a range of alternative approaches have attracted attention in recent years. This chapter discusses some of them, with emphasis on a group of probabilistic approaches that convert a given input into a probability distribution on set of REs (rather than into just one RE). I chose this focus because it raises questions that are not only pertinent to language production, but to many areas of human behaviour, because variation is (or appears to be) a feature of our behaviour in general.[1]

After discussing probabilistic approaches (sections 6.1 and 6.3), we turn to a family of approaches based on Constraint Satisfaction (section 6.4). As part of the same discussion we shall also introduce the generation of *relational* descriptions, such as "the book on the table", where the book is described *via* its relation to a table. Relational descriptions are of independent importance but they happen to have been addressed using Constraint Satisfaction first. Relational REs will feature again in section 6.5 and, finally, in chapter 10, where a more general solution to their generation will be proposed.

Having briefly discussed Constraint Satisfaction, we examine an approach to REG in which each property is associated with a *cost*, and which tends to use directed graphs as an underlying representation framework (section 6.5). The cost-based approach has inspired a fair amount of work in recent REG and offers an important generalization of the idea of a Preference Order. Finally, in section 6.6, we turn to a set of "dissident" approaches to REG, which beg to disagree with Dale and Reiter's research program (see section 4.3), insisting that reference should be understood as part of a larger communicative task.

1 Material related to the discussion of probabilistic approaches in this chapter can be found in conference papers such as [Gatt et al., 2011], [van Deemter et al., 2012c], [van Gompel et al., 2012], [van Gompel et al., 2014]; a comprehensive article on the PRO algorithm of section 6.3 has been submitted. Section 6.4 borrows some insights from [Krahmer and Van Deemter, 2012].

6.1 Variations in Language Production

Given almost any experimental study of language production, participants' responses vary substantially, even within a single experimental condition.[2] We saw an example in section 5.6, where some subjects were shown to be modelled more accurately by some algorithms, whereas others were modelled more accurately by others. Variation does not merely affect groups of speakers, but also the utterances of one and the same person: we call these two phenomena variation *between* and *within* speakers, respectively. Neither type of variations should come as a surprise, because they are reminiscent of what we know about other areas of behaviour. A seminal example comes from ballistic research around 1900, in which it was observed that the bullets of a skilled target shooter do not always hit the target, but pile up close to the bull's eye, with fewer and fewer strikes further away from it, giving rise to a bell-shaped probability distribution [Holden and van Orden, 2009].

The insight that within-speaker variation is an important factor in reference production was first made in [Gibbs and Orden, 2012], where variations in speakers' pragmatic choices are discussed, including the choice of whether or not to express privileged information [Horton and Keysar, 1996]. Holden proposed to explain these variations by assuming that "the bases of any particular utterance (...) are contingencies, which are to an under-appreciated extent the products of idiosyncracy in history, disposition, and situation" [Holden and van Orden, 2009]. Sociolinguists have long thought along similar lines, believing that within-speaker variation can be caused by language change and social register (e.g., dialects associated with different social strata); the idea is that different grammars are represented in the head of a single individual at the same time [Kroch, 2000]; for each utterance, the individual is thought to "choose" between different grammars, where the probability of choosing a given grammar is affected by the recent history of the individual; for example, who has she talked to recently?

At the time of writing, variations within and between people are starting to attract increasing attention throughout Cognitive Science, and there is a wide appreciation of the fact that variations in behaviour may have evolutionary advantages, because a more varied population is less likely to succumb to an environmental challenge. Yet there is no unanimity about how variability is

2 Variations in gestures are well attested too, see e.g., [de Ruiter et al., 2012].

best modelled and whether variation between and within people should be modelled by the same mechanism. For example, if a person's language production is influenced by recent utterances – as many researchers believe[3] – then this predicts both within-speaker and between-speaker variation. On the other hand, to the extent that variations between speakers are caused by differences in, for example, short-term memory (cf., [van Rij et al., 2013]), this explanation applies to variation *between* speakers only.

Although one could seek to model the precise effects of each factor that influences language production, in practice this may not be feasible. A small but growing number of linguists and psychologists believe that language may be best modelled probabilistically (e.g., [Chater and Manning, 2006]). Randomized algorithms have long been studied in pure computing science and mathematics, where they are valued for their speed and simplicity (e.g., see [Motwani and Raghavan, 1995] on Monte Carlo and Las Vegas algorithms). By *nondeterministic* algorithms we mean a generalization of this idea where probabilistic choices are not always purely random (i.e., fair). In the modelling of human behaviour, nondeterministic modelling is not a means to an end but the very goal of the modelling (e.g., [Lewandowsky and Farrell, 2011], chapter 4). Similar ideas have occasionally been explored in NLG [Belz, 2007].[4]

By looking at variation through the prism of a nondeterministic algorithm, it becomes possible to regard variation between and within speakers in the same way. In fact, we shall often think of the modelling of variation as one single task, without distinguishing between its two different flavours. Before presenting my own view of the matter, let's see how differences between speakers have been addressed in the past.

Many contributions to the TUNA challenges were data oriented, using corpus frequencies to find an algorithm that performs well. Sometimes this was done by hand, but sometimes, Machine Learning was employed to automatically construct a model. The learning algorithm may use features such as the number of distractors, the number of distractors having the same colour as the target referent, the number of objects having the same type, and so on. A number of researchers hit upon a natural extension of this idea: why not use

3 Some researchers even argue that retrieval of past instances of linguistic structure lies at the heart of all language processing [Bod, 1998], [Daelemans and van den Bosch, 2005]

4 Belz [Belz, 2007] compares a number of probability-based NLG systems, one of which, called greedy roulette-wheel generation, samples alternatives from a distribution, returning outputs in proportion to their likelihood. Belz observes that the corpus-based evaluation metrics on which her article concentrates do not do justice to this method.

the identity of the speaker as another feature? Versions of this approach were applied in [Bohnet, 2008], [Fabbrizio et al., 2008a, Fabbrizio et al., 2008b] (see our chapter 5, section 5.2), in [Viethen and Dale, 2010], and in [Mitchell et al., 2011b], [Mitchell, 2012]. We exemplify the issues by discussing the work of Viethen and Dale. Machine Learning is less popular in some corners of Cognitive Science (e.g., psychology) than others (e.g., Computational Linguistics), but it is a method that can coexist peacefully with other methods. This is because when Machine Learning finds rules or algorithms, this is not the end of the story: the rules can be tested in the same way as handcrafted rules or other hypotheses may be tested. Where this is done, it matters little whether the rules were found by Machine Learning or in some other way.

Viethen and Dale looked at two corpora containing RE elicited in settings showing scenes containing balls and cubes of different colours. The corpora are known as GRE3D3 and GRE3D7; the former involves three 3D objects, the latter seven. This makes the scenes even smaller than the TUNA domains, but their being 3D gives rise to some interesting referential possibilities: one object could be described as resting on another, for example. The authors used the C4.5 decision tree learning algorithm as implemented in the Weka workbench [Witten et al., 2011]. Applied in the usual fashion (and pruning decision trees to avoid overfitting), they learned two extremely simple decision trees. The one for GRE3D3 is representable as the following rule:

If some distractors have the same type as the target,
then use pattern R else use pattern D,

where pattern D contains just type and colour (e.g., "the blue ball) and R contains size as well (e.g., "the small red ball"). These rules, however, had very limited accuracy, defined as the number of instances predicted correctly divided by the total number of instances.

When Participant-ID (i.e., the identity of the speaker) was admitted as a feature, a very different picture resulted. When this feature was added to the other ones, the resulting participant-sensitive decision trees were much more complex (using the Participant-ID feature again and again) but had much better accuracy than before. In fact, Participant-ID was so powerful that even when used as the only feature (i.e., without using any features of the scene), the accuracy of the resulting decision trees was broadly comparable to the accuracy of the trees that were learned with all the other (i.e., nonparticipant) features.

Viethen and Dale's experiments show that the identity of the speaker matters greatly. Yet taken on its own, these investigations would be of little value for

predicting the behaviour of a speaker whose utterances have not been observed. Tantalizingly, Viethen and Dale suggest that future research may cluster speakers into groups, and even work with speaker profiles, aiming to construct separate algorithms for each profile. If, in the future, profiles allowed one to automatically classify a new speaker (maybe based on a few test utterances, or based on other characteristics of the person), then this would mean a substantial leap in our understanding of the differences between human speakers.

Unfortunately, this approach is not easily extendible to within-speaker variation (unless this could be linked to the moods of a speaker, essentially treating a person as a conglomerate of personalities). Moreover, the rules that are learned are fully deterministic, producing the same output always for a given input. It is time for us to look at approaches to REG that have probability at their heart.

6.2 Bayesian Models of Reference

Bayesian models are thought to be suitable for modelling nondeterministic mental processes, an idea that is sometimes referred to via the phrase *probabilistic brain* [Pouget et al., 2013]. I will discuss here a model of reference production by Frank and Goodman, which can be positioned within this tradition [Frank and Goodman, 2012], and I will suggest ways in which the model may be modified to bring it in line with other experimental results.

Frank and Goodman's model, which looks at comprehension and production in tandem, assumes that language users are rational in the following sense: speakers refer by choosing a property that has a high Discriminatory Power; hearers comprehend an RE by choosing a referent that is probable given the RE, maximizing $P(r|w, C)$, the probability that a word w, uttered in context C, denotes the referent r. Using standard Bayesian reasoning – which often seeks to reverse the "direction" of a conditional probability – the idea is that hearers accomplish this task by maximizing

$$\frac{P(w|r, C) * P(r, C)}{\sum_{r' \epsilon C} P(w|r', C) * P(r')}. \tag{6.1}$$

The denominator is a normalizing constant that compares r to all the objects r' in the domain. $P(w|r, C)$ is the *likelihood* of the word w being chosen to refer to r in context C. The likelihood term interests us particularly, because it concerns reference *production*, representing the probability that w is chosen to refer to the referent r (see below). $P(r, C)$ is the *prior* probability that r is the referent, estimated by asking subjects, in a TUNA-like setting involving just

three objects, "Imagine someone is talking to you and uses a word you don't know to refer to one of these objects. Which object are they talking about?" $P(r, C)$ can be thought of as estimating the salience of r in the context C.

The likelihood term, $P(w|r, C)$, is estimated using the Discriminatory Power of w, much like the Greedy Algorithm of chapter 4. The prediction is that speakers tend to choose words that rule out large numbers of distractors. This is modelled as a probability distribution over words; comprehension is modelled as a distribution over REs. Probabilities of this kind allow two interpretations: a *knot cutting* approach says that the RE with the highest probability should always be chosen; a *nondeterministic* interpretation says that each output with non-zero probability may be chosen, in accordance with the probability assigned to it by the model. I will assume that the authors had the latter interpretation in mind.

The models were tested with human participants who saw a domain of three objects. To test the *comprehension* model, participants were asked to bet on each of the three objects, with the total of each participant's three bets in a given situation summing to 100 points (e.g., 80 for one object and 10 for each of the other two), resulting in what can be interpreted as a probability distribution over the three referents (e.g., 0.8, 0.1, 0.1). Similarly, when they tested the *production* model, the authors asked participants to bet on a word, resulting in a probability distribution over words. Good correlations between predictions and observations were reported in both cases.

The authors see themselves as modelling "a referential communication game", and it is by means of (betting) games that the model was tested [Frank and Goodman, 2012]. Although the paper "synthesizes and extends work on (...) systems for generating referrIng expressions", it only deals with extremely simple REs, which contain only one property. Moreover, the communication game does not involve natural language; actual English NPs do not play a role. Perhaps unsurprisingly given its exclusive emphasis on Discriminatory Power (with no place for Intrinsic Preference), experiments suggest that the production side of the model is inadequate as a predictor of people's production of REs. My colleagues and I confronted human speakers with situations that closely resemble Frank and Goodman's, asking them to produce an RE. In situations where colour and size were equally discriminating (see Figure 6.1), for example, their model predicts that size and colour are chosen with the same likelihood, but we found that speakers used REs containing only colour on almost 80% of cases, size only on fewer than 4% of cases, and both colour and size in the remaining 17% [Gatt et al., 2013a] [Gatt et al., 2013b]. These

Figure 6.1

A domain from [Gatt et al., 2013b]. In this case the referent can be identified using either colour ("green") or size ("large").

data suggest that a Bayesian approach based on Discriminatory Power alone does not work well.

Frank and Goodman's work can be seen in a more generous light if we ignore their emphasis on Discriminatory Power, focussing on the broad architecture of their proposal, which elegantly combines production and comprehension into one model. Intriguingly, it predicts that hearers who calculate the probability $P(r|w, C)$ take $P(w|r, C)$ into account in this calculation.

For example, suppose a domain contains three equally salient men, a, b, and c. Suppose (only) b and c have a moustache, whereas (only) c wears glasses. The model predicts that hearers tend to understand the word "(the man with a) moustache" as referring to b rather than c, because c is more likely to be called "(the man with) glasses", because "glasses" has higher Discriminatory Power than "moustache" in this situation. The implication is that it would be *rational* for speakers to deviate from the production model based on Discriminatory Power and refer to b saying "the moustache" even though, strictly speaking, this leaves one distractor, c, to be removed.[5]

A promising model results if we turn Frank and Goodman's model upside-down, using Bayesian reasoning to derive $P(w|r, C)$ from probabilities that it might be easier to estimate. For example, one could use the following version of Bayes' Law,

$$P(w|r, C) = \frac{P(r|w, C) * P(w, C)}{P(r, C)}, \tag{6.2}$$

interpreting $P(r|w, C)$ as the probability that a word w, uttered in context C, denotes the referent r; this probability could be influenced not just by the

5 This type of reasoning, in which hearers and speakers take each other's point of view into account, is familiar from bi-directional Optimality Theory, see e.g., [Blutner, 2000].

salience of r within C, but also by how typical a referent r is for w. The term $P(w, C)$ could be interpreted as the degree of preference of w, in the manner of the Incremental Algorithm (or the cost-based approach of section 6.5). In keeping with the spirit of Bayesian reasoning [Howson and Urbach, 1996], various resources can be brought to bear on estimating each term including, for example, insights from Prototype Theory, which say that properties are true of objects to different degrees [Rosch, 1978], see our section 3.4 and elsewhere). For example, a swallow is a more prototypical example of a bird than a duck, so if r_s is a swallow and r_d is a duck, then $P(r_s|bird) > P(r_d|bird)$. In the same spirit, corpus frequencies could be employed to estimate $P(w)$ (recall sections 5.4 and 5.5, where corpus frequencies were employed to determine the Preference Order of an Incremental Algorithm).

Algorithm 8 Sketch of a possible new Bayesian approach to REG

Input: A communication context C, which determines a domain containing a referent r and a non-empty set of distractors. A candidate word w for describing r in the context C, and information about the following: Rosch-style prototypicality; the degree of preference associated with w; and the salience weight of r.
Output: A predicted probability $P(w|r, C)$ that w is chosen for describing r in C.

1: Use the degree of prototypicality of r for w to estimate $P(r|w, C)$
2: Use the degree of preference of w to estimate $P(w, C)$
3: Use the salience weight of r to estimate $P(r, C)$
4: Apply Bayes' Law to the outcomes of (1)-(3) to compute $P(w|r, C)$

As it stands, the new Bayesian production model inherits Frank and Goodman's limitation to REs containing just one word; further modifications would be needed to turn it into a complete REG model.

6.3 Probabilistic Referential Overspecification: the PRO Algorithm

Instead of designing a novel approach to REG, it would be possible to keep what is good about the classic REG algorithms, while somehow adding variation. One possibility, proposed in [van Deemter et al., 2012c] and tested in [van Gompel et al., 2012, Gatt et al., 2013b, van Gompel et al., 2014], is to turn current deterministic algorithms into nondeterministic ones. Consider the Incremental Algorithm once again (chapter 4). The original, deterministic, IA always generates the same RE in given situation. Yet even Pechmann's early data (which motived the IA) are not covered well by this approach.

For example, in situations where Pechmann asked attention for overspecification, *minimal* descriptions were produced on as many as one quarter of trials [Pechmann, 1989]. One way to account for this type of variation would be to use a nondeterministic approach that lets the Incremental Algorithm check properties in different orders, each of which has a particular probability. Such an approach would not directly tell us *why* these probabilities apply, but it could plausibly be argued that frequencies in the language use to which speakers have been exposed in the past may have something to do with it.

For concreteness, suppose when referring to a small black cup in the context of a large white cup and a large red cup (so colour or size can be used to uniquely characterize the target), speakers produce *the black cup* four times more often than *the small cup*. In this case one might conclude there is a 80-20% colour-size preference. A nondeterministic Incremental Algorithm could explain this pattern by positing that speakers first check colour in 80% of cases, but size in the remaining 20% (and the category *cup* is added to make sure that the RE can be expressed by means of a noun phrase).

Having determined the colour-over-size preference, one can predict how often overspecification occurs in other situations. When referring to a small black cup in the context of a large white cup and a large black cup, the algorithm initially chooses colour over size in 80% of cases, but because this does not uniquely identify the target, it subsequently adds size, resulting in an overspecified expression (*the small black cup*) in 80% of cases. In the other 20%, it first checks size, and because this uniquely identifies the target, colour will not be added. An 80-20% split is also predicted to occur when the same target (a small black cup) occurs in a context with a small white cup and a large white cup (so colour is required). In 80% of cases, colour is checked first, and because this uniquely identifies the target, the algorithm produces *the black cup*. In the other 20%, size is selected first, but because it does not uniquely identify the target, colour is added, resulting in *the small black cup*. Thus, the Nondeterministic Incremental Algorithm makes clear quantitative predictions that arise from the fact that colour is usually checked before size. Other algorithms can be made nondeterministic in similar ways.

Existing metrics of humanlikeness (see chapter 5) were not designed to measure the extent to which an algorithm reflects the variation in a corpus. This is as true for within-speaker variation as it is for between-speaker variation [van Deemter et al., 2012c]. Consider an example involving just one reference task, to which a speaker is exposed on two occasions: on occasion *a* she utters NP_1; on occasion *b* she utters NP_2. Suppose these NPs are as different as they

(a) Colour suffices for identifying the referent

(b) Size suffices for identifying the referent

(c) Colour or size suffices for identifying the referent

Figure 6.2

Three types of domains on which the PRO *algorithm was tested. In each case, the candle in the middle is the intended referent.*

can be, so if a generated RE has a Dice score of 1 for one, it has a score of 0 for the other. Now consider two algorithms, each of which is run twice. One algorithm is nondeterministic and generates the two human-produced NPs, but the other behaves deterministically, generating NP_1 on both runs:

Occasion *a*: NP_1. Occasion *b*: NP_2
Algorithm 1: NP_1 (first run); NP_2 (second run)
Algorithm 2: NP_1 (first run); NP_1 (second run)

Algorithm 1 captures the variation among the two occasions much better than Algorithm 2 (which does not show any variation). Existing metrics, however, attribute the same score to both algorithms, because these metrics compute the extent to which the Logical Forms generated by a given algorithm match the information contents of human-produced descriptions *on average*, comparing each generated description with each human-produced one. Because both descriptions, NP_1 and NP_2, match one human-generated description fully (leading to a score of 1) while failing to match the other one entirely (scoring 0), both algorithms end up with the same averaged score of 0.5.

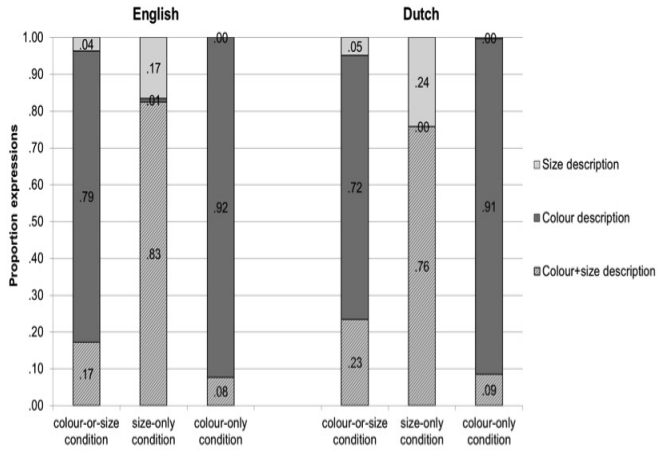


Figure 6.3

Proportions of each of the three possible types of REs*, in each of the three conditions (colour, size, colour or size) and in both Dutch and English.*

Trying out these ideas on simple domains that contained just three objects (Figure 6.2), Roger van Gompel, Albert Gatt, Emiel Krahmer, and I found that algorithmic models can be constructed that match variations in human referential behaviour very closely [Gatt et al., 2011, van Gompel et al., 2012, Gatt et al., 2013b, van Gompel et al., 2014]. References were elicited in both Dutch and English under three different conditions (Figure 6.2): (1) the colour (but not the size) of an object suffices to distinguish the referent; (2) the size (but not the colour) suffices to distinguish it; and (3) both the colour suffices and the size suffices (we call this the colour-or-size condition). The original Incremental Algorithm, with a Preference Order in which colour precedes size, makes the following predictions, as one can easily verify:

Condition 1 (colour suffices): select colour only (100% of cases).
Condition 2 (size suffices): select colour and size (100% of cases).
Condition 3 (colour suffices and size suffices): select colour only (100% of cases).

For example, in Condition 2, colour is selected first, because colour is the most highly preferred attribute and colour does remove a distractor; however, the referent has not yet been identified, therefore size needs to be added. Human-generated REs showed a different picture. Figure 6.3 shows that the preference for colour over size is far from absolute, in both Dutch and English.

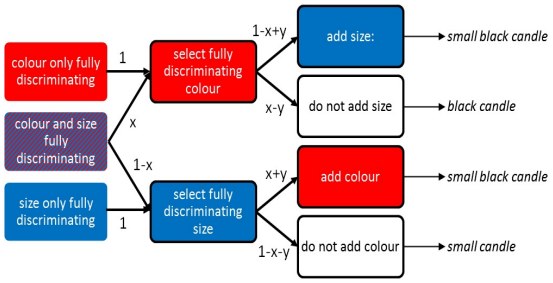


Figure 6.4

The PRO *model applied to a small domain.*

To model the data, we used the parameter x to denote the probability of selecting colour as opposed to size. Given that the Preference Order of old is now replaced with probabilities, let us call x the *preference degree* of colour. We used y as a parameter that governs the likelihood of adding a second property in a situation where the first property was already distinguishing; in practice y tends to be negative, so it is not itself a probability but a parameter that is employed to change a given probability. Consider the colour-or-size condition, for example. Figure 6.4 shows that the model, called Probabilistic Referential overspecification (PRO), starts by choosing between colour (with probability x) and size (with probability $1 - x$). Suppose it chooses colour. Next, the algorithm chooses between adding size (with probability $1 - x + y$, causing overspecification) and adding no further property (with probability $x - y$). Thus, the probability of choosing only colour, in this condition, equals $x.(x - y)$. Note that the other possibility, of using both colour and size in this condition, can be reached via two different paths, hence its probability is $(x.(1 - x + y)) + (1 - x).(x + y)$.

Crucially, the values of the parameters x and y were based on held-out data: their value, in a given condition, is determined by looking at the REs obtained in the two other conditions. For example, the values of x and y in the *colour-only* condition were predicted using the best-fitting values of x and y in the *size-only* and *colour-and-size* conditions. In this way, training and test sets were kept separate, and overfitting was avoided. By training on conditions that differ from the test condition, this approach is arguably more thorough than what is usually seen in Machine Learning. It might be argued that our procedure was biassed in favour of PRO, because the algorithm was trained on data

that resemble the test data (because the three conditions were somewhat similar). However, all the other algorithms with which PRO was compared were given the same advantage, so this bias should apply to all algorithms.

Using values for the parameters x and y that were optimal for the training corpus, we arrived at a nondeterministic procedure that, faced with the domains described above, produced a distribution of RE types that resembled very closely the distribution that exists in human-produced REs. Evaluation on the test corpus showed the resulting model to easily outperform its competitors, including the Bayesian model of section 6.1 and a probabilistic version of the Incremental Algorithm. Evaluation was performed using the Bayesian Information Criterion (BIC), a standard approach in computational modelling (e.g., [Lewandowsky and Farrell, 2011], section 5.4). The idea is to compute, *given* a particular probabilistic model, the probability that the data are exactly as observed.[6] This basic idea is moderated by a mechanism that discounts the number of free parameters of the model, a procedure sometimes known as a Bayesian Occam's Razor.

In a later experiment with Van Gompel, Gatt, and Krahmer, the same approach to data modelling was applied to REs produced under two different conditions, namely, a condition under which the size contrast was minor (as in Figure 6.2) and one in which this contrast was much greater. [van Gompel et al., 2014] As expected, the two conditions gave rise to very different preferences. In particular, when the size difference was large, PRO used size much more often than before. This application of the training method illustrates handsomely what sorts of insights it can generate, improving our understanding of the idea of Intrinsic Preference, which plays such a key role in REG algorithms (see e.g., section 3.4).

Extrapolating from the PRO model of figure 6.4, and taking additional experiments into account, a full REG model emerged, which we will call the PRO *algorithm*. The algorithm, which follows a broadly monotonic structure that resembles the classic REG algorithms to some extent, uses a (simple) dedicated mechanism for overspecification, and guarantees that if a fully discriminating property exists (i.e., a property that, by itself, removes all distractors), then one such property is selected. PRO (see Algorithm 9 below) extrapolates from our

6 Suppose, for example, your model asserts that a coin is fair. If you throw the coin twice, then, assuming the model, the probability of throwing heads and tails equally often (i.e., once each) equals $1/2$ (i.e., 2 out of 4 possible outcomes). If you throw the coin four times, this probability (i.e., throwing each of head and tail twice) is only $3/8$ (i.e., 6 out of 16 outcomes), and so on.

experiments in [Gatt et al., 2011, van Gompel et al., 2012, Gatt et al., 2013b] by iterating through a number of steps, each of which either adds a property or not. As in other monotonic algorithms, no attribute is ever withdrawn.

As usual, the pseudo-code below will focus on concrete properties, rather than Attributes (which are essentially clusters of similar properties). This will have three advantages. First, it simplifies the statement of the algorithm. Second, it prevents the problems with overlapping values that we encountered in section 4.7. Finally, it opens the door to treating the different values of an Attribute differently, which may be desirable in light of section 3.4 (see also section 16.2) – assuming that, unlike in the PRO model above, the parameters x and y receive values *per* property, not *per* Attribute.

The pseudo-code below does not mention the TYPE of the referent because it assumes that TYPE is always a part of the RE. As before, r is the target referent. $\mathcal{P}$ is the set of all available properties that hold true of r; $\mathcal{P}$ shrinks as properties are added to the Logical Form $\mathcal{D}$ that is generated. M is the set of remaining domain objects, which likewise gets smaller. Updating does what we have come to expect: suppose a property P has been chosen, then $\mathcal{D}$ is updated by adding P as an element of $\mathcal{D}$; $\mathcal{P}$ is updated by removing P from $\mathcal{P}$; and M is updated by intersecting M with the extension $[\![P]\!]$. First the aim is to identify the referent (lines 1-10); overspecification is addressed after that (lines 12-14). Line 5 ensures that a property is only added to $\mathcal{D}$ if, by doing so, one or more distractors are removed from M.

PRO uses nondeterministic choice a number of times. Twice this is done by $\mathcal{P}$-*choose*, which chooses a property with a probability that equals its Preference degree. It is used when several properties remove all distractors and when looking for a new property in a situation in which the Logical Form $\mathcal{D}$ is not a distinguishing description. Probabilistic choice at line 12 is governed by a subtly different device called $\mathcal{P}$-*choose*$^+$, a modification of $\mathcal{P}$-*choose* that chooses not only between the properties remaining in $\mathcal{P}$ (in accordance with the Preference Degree of the property) but also between adding a property and adding nothing (*Stop*), in which case the construction of $\mathcal{D}$ is terminated. If a property is chosen, it is added to $\mathcal{D}$, causing overspecification. $\mathcal{P}$-*choose*$^+$ can fire any number of times, until *Stop* is chosen. Overspecifications of any

length are possible, but the greater the number of overspecifying properties in a description is, the smaller the probability of this description is.[7]

Algorithm 9 Probabilistic Referential Overspecification (PRO)

Input: A domain containing a referent r and a non-empty set of distractors M. A set $\mathcal{P}$ of properties true of r. Functions $\mathcal{P}$-*choose* and $\mathcal{P}$-*choose*$^+$, as in the text.
Output: A distinguishing description $\mathcal{D}$ of r. $\mathcal{D}$ is chosen probabilistically.

1: Start out with an empty $\mathcal{D}$
2: **if** there exists a property in $\mathcal{P}$ that removes all distractors (on its own) **then**
3: $\mathcal{P}$-*choose* one such property, for example P
4: $D := \{P\}$
5: **while** True **do**
6: Remove from $\mathcal{P}$ all properties that are true of all elements in M
7: **if** $\mathcal{P} = \emptyset$ **then**
8: **return** $\mathcal{D}$
9: **else if** $\mathcal{D}$ is not distinguishing yet **then**
10: $\mathcal{P}$-*choose* a property from $\mathcal{P}$
11: Update $\mathcal{D}$, $\mathcal{P}$, and M
12: **else**
13: $\mathcal{P}$-*choose*$^+$ between *Stop* and the properties remaining in $\mathcal{P}$
14: **if** one of the properties in $\mathcal{P}$ is chosen **then**
15: Update $\mathcal{D}$, $\mathcal{P}$, and M
16: **else**
17: **return** $\mathcal{D}$

In the articles cited above, we describe an experimental comparison between the PRO model and its main competitors, whose outcomes suggest that PRO's combination of Preference Degrees and Discriminatory Power (as in lines 1-3 of PRO) is on the right track. Yet, there may be room for improvement, especially in complex referential situations. For instance, Discriminatory Power gets only one chance to make itself felt: after line 4, it plays no role. Maybe Discriminatory Power should get a second chance if there exists a property that removes all distractors once the first property has entered the Logical Form; only further experimentation can tell. Likewise, there is no experimental evidence concerning repeated overspecification (given that $\mathcal{P}$-*choose*$^+$ is embedded inside the *while True* loop), so the fact that the algorithm permits any amount of overspecification is only a guess.

7 This is because an overspecified description of length $n + 1$ is constructed by building a description of length n first, then adding a further property, using $\mathcal{P}$-*choose*$^+$. This addition tends to happen with a probability far less than 1, unless y is very large.

Although some of these issues may be resolved, it is inevitable that some untested predictions are made. This is part and parcel of what it means to construct a model that covers unobserved situations (involving domains that contain any number of objects, with any number of attributes). Computational cognitive models can be general enough to make decisions in all possible cases, but they cannot be complete unless they jump to conclusions sometimes.

Results of a similar nature to the ones that motivate PRO were obtained in connection with the nondeterministic algorithm of Mitchell and colleagues [Mitchell et al., 2013c], which was evaluated on both the TUNA and the GRE3D3 corpus. Another recent contribution to the debate is [FitzGerald et al., 2013], where distributions of more complex types of REs (i.e., Boolean REs, see chapter 8) are obtained using a statistical method called density estimation. Although all three models were obtained in settings that emphasize *between*-speaker variation, they lend themselves well for modelling *within*-speaker variation too. It may be too early to choose one nondeterministic algorithm over another, but it does appear that nondeterminism offers an attractive range of possibilities for the modelling of within-speaker variation.

On a speculative note, within-person variation might one day lead to drastically new models such as, for example, quantum models of human reasoning [Bruza et al., 2009]. Quantum models permit *superposition* states, originally invented for theoretical physics: "To be in a superposition state means that all possible definite values (...) have potential for being expressed at each moment" [Wang et al., 2013]. For instance, a speaker might be in a superposition state that is ambivalent between an inclination towards a low amount and an inclination to a high amount of overspecification. The benefits and drawbacks of quantum models are not clear to me yet, but it does seem that behavioural variation offers scope for exciting new computational models.

6.4 Constraint Satisfaction for REG

We now move away from variations in language production and the models that aim to justice to them, turning to two computational paradigms that differ from the mainstream of REG and that contain lessons about reference and the manner in which REs may be produced. We start with the paradigm of Constraint Satisfaction. We shall see that Constraint Satisfaction uses essentially the same algorithm always: what is different, from one problem to the next, is

the information that represents the problem. For this reason, the present section will not show any algorithmic pseudo-code.

Constraint Satisfaction arose as a computational paradigm that allows efficient solving of computationally intensive combinatoric problems [van Hentenryck, 1989]. It allows an elegant separation between the declarative specification of a problem (e.g., the problem of referring to a given individual in a given context) and the details of its solution (e.g., a particular REG algorithm). Moreover, the approach has been shown to allow remarkably fast solutions. Technically a *constraint satisfaction problem* involves variables, each of which is associated with a range of permitted values and a set of *constraints*. Solving the problem means finding an assignment of values to each variable that is *legal*, meaning that each value is within the range associated with the variable while, crucially, the assignment meets all constraints. Constraint Satisfaction involves searching through the space of all possible assignments of values to variables and comes into its own when there are dependencies between variables [Kumar, 1992].

An example that is often given to illustrate the power of Constraint Satisfaction is the "eight queens" problem: find a way to position eight queens on a chess board in such a way that no two queens can capture each other. Suppose we analyse this problem using 8 variables, $a, ..., h$, each of which stands for a column of the chess board, and each of which has $1, ..., 8$ as permitted values, depending on which of the eight squares in this column hosts a queen; more than one is impossible given that these would capture each other. The constraints say that none of the queens on the board can capture each other (i.e., share a column, row, or diagonal). Suppose we start our search by assigning the value 1 to the variable a, tentatively placing the first queen on the first square of the a column. Turning to the b column, in order to place the second queen, only the values $3, ..., 8$ need to be considered, because squares 1 and 2 are under attack from the queen at $a1$. If we decide to place this second queen on $b3$, then the set of permitted values has shrunk to just four values: $5, ..., 8$. Pruning the search tree in this manner is an example of the kind of computational economy afforded by Constraint Satisfaction.

Relational descriptions. Constraint Satisfaction was one of the first frameworks proposed for REG [Dale and Haddock, 1991], where it tackles a type of RE that has so far been neglected in this book, and which are known as *relational* descriptions. A relational description refers to a referent r *via* an entity other than r. For example, "the cup on the table" is a relational description

because although its main purpose is to refer to a cup, it contains another NP, "the table", which helps to identify the cup. Given that this can go on indefinitely ("the cup on the table in the corner of the ..."), one might speak of *recursive* descriptions.

The construction of simple relational Logical Forms can be seen as a Constraint Satisfaction Problem in different ways. Dale and Haddock proposed an approach in which predications (like "being a cup", "being on top of ...") are constraints that can apply to each domain object, so REG is the search for a set of constraints that jointly identify a given target referent. An alternative approach in [Gardent, 2002] uses a variable V, which takes as values sets of predications, where a predication can be a property $P(x)$ (saying that x has the property P) or a relation $R(x, y)$ (saying that x stands in the relation R to y). If r is the target referent, solving the Constraint Satisfaction Problem means assigning a value to V in accordance with two constraints, namely: (1) all predications in V are true of r, and (2) for each distractor d there is a property in V that is false of d. In section 6.4, these ideas will be developed further.

Dale and Haddock's article made a number of important observations: First, it is possible to identify an object through its relations to other objects without identifying each of these objects. Consider a situation involving two cups and two tables, where one cup is on one of the tables. In this situation, neither "the cup" nor "the table" is distinguishing, but "the cup on the table" succeeds in identifying one of the two cups. Second, descriptions of this kind can have any level of depth: in a complex situation, one might say "the white cup on the red table in the kitchen", and so on. To be avoided, however, are the kinds of repetitions that can arise from descriptive *loops*, because these do not add information. It would, for example, be useless to describe a cup as "the cup to the left of the saucer to the right of the cup to the left of the saucer . . . ".

The proposed handling of relations has a number of important strengths. The approach does not run into the aforementioned descriptive loops, because a set of properties (being a set) cannot contain duplicates. Moreover, it generalizes effortlessly to logically more complex situations, involving negation, sets, and relations with arbitrary numbers of arguments (as in "the present that John gave to Harry"). This does not mean, however, that it is the last word on the generation of relational descriptions, for Constraint Satisfaction is compatible with many different search regimes (e.g., [Russell and Norvig, 2003], chapter 5; [Kumar, 1992]), which enables this approach to emulate many of the algorithms discussed in the previous chapters. What generation algorithms

suit relational descriptions best is still an open question. As Emiel Krahmer and I wrote in our survey in 2012:

> Various researchers have attempted to extend the IA by allowing relational descriptions [Horacek, 1996, Krahmer and Theune, 2002, Kelleher and Kruijff, 2006a], often based on the assumption that relational properties (like "x is on y") are less preferred than non-relational ones (like "x is white"). (...) It seems, however, that these attempts were only partly successful. One of the basic problems is that relational descriptions (...) do not seem to fit in well with an incremental generation strategy. In addition, it is far from clear that relational properties are always less preferred than non-relational ones [Viethen and Dale, 2008]. (...) On balance, it appears that the place of relations in reference is only partly understood, with much of the iceberg still under water. [Krahmer and Van Deemter, 2012]

Constraint Satisfaction has been applied to REG a number of times in recent years. Two strands of work are worth singling out. One is Claire Gardent's work on the generation of references to sets. The other is a line of work by Matthew Stone and colleagues, which seeks to connect REG with a wider class of issues in the pragmatics of natural language. We start with the former.

Reference to sets. Constraint Satisfaction has also been used in connection with reference to sets [Stone, 2000], a topic discussed more fully in chapter 8. Here we focus on [Gardent, 2002], which aimed to generate REs that refer to a set of objects distributively, that is, by means of properties that hold true of each of the elements of the set but are false of everything else. In its simplest form, the semantic content of an RE for a target set S is formalized as just one variable P, which ranges over sets of properties. The challenge is to find suitable values (i.e., sets of properties) for P. To be "suitable", values need to fulfil two REG-style constraints:

1. All the properties in P are true of all elements in S.
2. For each distractor d there is a property in P that is false of d.

This approach can be extended to address a variety of harder problems. Matthew Stone, for example, applied a similar approach to the generation of *collective* references, as when we say "the parallel lines at the top of the screen" to refer to a set of lines that run parallel *to each other*. Stone showed

that such logically more complex RE can be generated if P is permitted to contain second-order properties, which are properties that hold true of a set (e.g., a set of two lines).

Gardent herself sought out another problem: following the work that will be discussed in chapter 8, she allowed that properties may be used negatively, as when we say, speaking of a chess board for instance, "the squares that are not occupied". The content of an RE is now formalized as a pair of variables:

$$\langle P_S^+, P_S^- \rangle.$$

The first variable ranges over sets of properties that are *true* of the elements in S and the other over properties that are *false* of the elements in S. The aim is to find values for these variables, subject to three constraints:

1. All the properties in P_S^+ are true of all elements in S.
2. All the properties in P_S^- are false of all elements in S.
3. For each distractor d there is a property in P_S^+ that is false of d, or there is a property in P_S^- that is true of d.

The third clause says that every distractor is ruled out by either a positive property (i.e., a property in P_S^+) or a negative property (i.e., a property in P_S^-), or both. The approach is further extended to accommodate disjunctive properties (as in "the squares that are occupied by a rook or a bishop") [Gardent, 2002].

Gardent opts for a *propagate-and-distribute* strategy, first looking for single properties, next for combinations of two properties, etc., increasing the logical complexity of the RE stepwise. This amounts to a Full Brevity algorithm. The "propagating" efficiencies of Constraint Satisfaction programming are exploited, for example, by making sure that properties in P_S^+ are no longer considered for inclusion in P_S^-.

The algorithms proposed in [Gardent, 2002] always yield a minimally distinguishing Logical Form for a target, provided one exists. In view of our earlier discussions this may not be the best choice. However, as the discussion at the end of her paper rightly suggests, her algorithm could be adapted to accommodate very different REG algorithms. Constraint Satisfaction does not force one to opt for Full Brevity any more than the other approaches in this chapter: its essence does not lie in brevity but in permitting a declarative statement of the REG problem, and in the use of efficient search strategies.

Summing up, Constraint Satisfaction offers an elegant and powerful approach to REG that is perhaps particularly suitable for tackling issues that

go beyond the classic REG problem. It is not clear, however, that the method can be generalized to the types of REs that will be discussed in chapter 10, which call for the power of predicate-logical deduction, as we shall argue.

6.5 Krahmer et al.'s Cost-Based Approach

A relatively recent approach to REG is commonly known as the graph-based approach [Krahmer et al., 2003]. Graphs are popular in Artificial Intelligence, and fast algorithms for dealing them are widely available. However, I shall argue that the use of graphs is a less important feature of Krahmer and colleagues' approach to REG than the idea that the inclusion of a given property in an RE comes at a certain precisely quantifiable *cost*, and that this cost may be used to steer the generation process. It is possible to assume that all properties associated with the same attribute have the same cost (cf., section 4.6), but it is not necessary. Likewise, it would be possible to use a variant of the *monotonic* approach to REG (section 4.5), adding properties one by one, always choosing a property whose cost is least, but this is not what the authors proposed.

A graph-based definition of reference. The graph-based approach starts with the idea that a referential domain and an RE can each be modelled as a labelled directed graph: the Scene Graph and the Description Graph, respectively. In both graphs, nodes (vertices) represent individuals in the domain of discourse, whereas properties and relations are modelled as edges (arrows); a label attached to an edge says what properties or relation this edge represents.

Before discussing the idea that REG can be defined as a comparison between a Scene Graph and a Description Graph, we need to explain how graphs can express atomic information. Figure 6.5 from [Krahmer and Van Deemter, 2012] shows a Scene Graph involving two men and one woman. Properties such as "man" are modelled as loops (edges beginning and ending in the same node); relations such as "left of" are edges between nodes. Given a Scene Graph and a Description Graph containing a node r (the target referent), one asks whether the Description Graph can be "placed over" r in the Scene Graph, in which case the Description Graph is consistent with what the Scene Graph says about r. If this is the case, the crucial question is whether there are any other vertices over which the Description Graph can also be placed. If not, then the Description Graph refers uniquely to r given the Scene Graph [van Deemter and Krahmer, 2007].

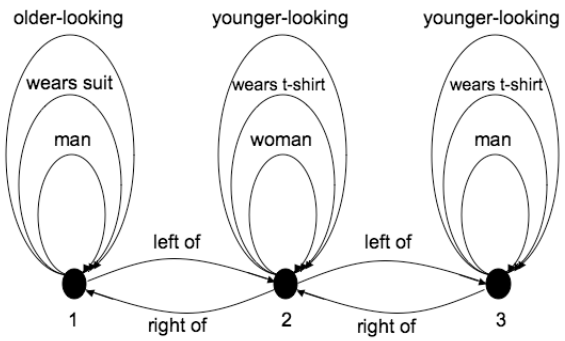


Figure 6.5

A Scene Graph: a domain represented as a labelled directed graph.

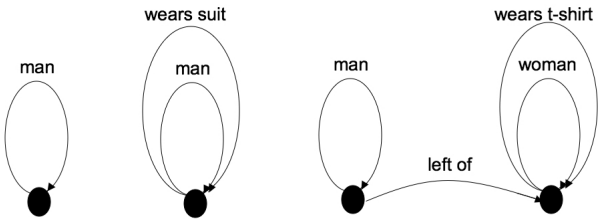


Figure 6.6

Three description graphs that refer to node 1 *in Figure 6.5. Two of the three refer to node 1 uniquely.*

Figure 6.6 shows a number of referring graphs that can be placed over our target referent d_1. The pair consisting of the leftmost description graph and its only node fails to distinguish the target in the Scene Graph, because it can be "placed over" the Scene Graph in two different ways (over nodes 1 and 3).

Using graphs to compute a referring expression. The graph-based approach can implement a variety of algorithms. In practice, algorithms have focussed on finding the lowest-cost Description Graph that refers uniquely to the intended referent. The algorithm of [Krahmer et al., 2003] uses "branch and bound", a fast search strategy familiar from a range of optimization problems. The original algorithm constructs Description Graphs that are subgraphs of the Scene Graph, and which contain the target referent, starting with just one node. New

edges and vertices are added recursively, removing distractors in the process. In its search for a minimal-cost RE, the algorithm stores the lowest-cost RE that it has found so far in a variable called *bestgraph*; if and when an RE with a lower cost is found, the current best graph is "dethroned". Details of the algorithm are principally motivated by speed [Krahmer et al., 2003].

Strengths and limits. By minimizing the cost of an RE, this algorithm generalizes the idea of minimizing the length of an RE. The two ideas coincide if all properties have the same cost, but cost offers possibilities that length does not. First, as observed in [Krahmer et al., 2003], the cost of a property can depend on a variety of factors, including Intrinsic Preference: if more highly preferred attributes have lower cost, then the algorithm resembles the Incremental Algorithm; an alternative is to determine cost *via* frequencies in a corpus [Theune et al., 2011]. Furthermore, cost is more fine grained than Intrinsic Preference, because it replaces the mathematical concept of a linear order (i.e., one attribute being preferred over another) with a metric that tells us *how much* more preferred one attribute is than another. This also makes it possible to assign the same cost to two properties, or to assign a nil cost to a property, to facilitate certain types of overspecification. Minimizing cost, in other words, is a more flexible idea than minimizing length.

The main ideas discussed here could have been implemented without graphs: most frameworks that permit the expression of properties and relations permit the assignment of costs as well, and the use of search strategies such as *branch and bound*. Conversely, a graph-based framework does not enforce the use of costs or a particular search strategy: it would be a simple matter to implement the classic REG algorithms directly (i.e., without using costs) using graph-based knowledge representations.

Nonetheless, the ability to control both the cost of each property and the order in which they are tried [Viethen et al., 2008] makes the graph-based approach potentially powerful. Excellent results were obtained when variants of the above-described algorithm were submitted to the TUNA evaluation challenges (see chapter 5) of 2007 and 2008; the best results were obtained using a cost function based on corpus frequencies where some properties had a cost of 0, and where properties were tried from cheapest to more expensive.[8]

8 The average Dice score of this algorithm was .71 for the furniture corpus and .67 for the people corpus [Theune et al., 2007]. Later tests focussing on relational descriptions (which did not play a role in the TUNA challenges) suggest good performance in that area as well.

To see why it matters in which order edges are tried, suppose attributes such as TYPE and COLOUR are assigned a cost of 0 (as in [Krahmer et al., 2008]). Then consider a Description Graph G that does not contain colour and refers uniquely to a referent r. Now compare the current best graph G with the graph G' that results from adding to G an edge representing the colour of the referent. Under the assumptions stated, G and G' have the same cost, so G' does not de-throne G once G has become the current best graph; consequently, the RE generated will not contain colour. If, however, the algorithm had started with colour, then an RE would have been generated that does represent colour. In other words, cost does not always completely replace preference. Henceforth, we shall assume that Intrinsic Preference, in the style of Dale and Reiter, can be modified to have the same benefits as Krahmer's cost, so attributes (or properties) are associated with a numerical *degree* of preference.

Inherent in this approach is the idea that cost is bad, and that the production of a distinguishing description is all that counts. Although, as we have seen, these ideas do not preclude over-generation, they are at odds with the idea of adding overspecifying properties in order to help the reader (section 12). To do this, an algorithm would have to compare the trouble of adding a property with any benefits that this brings for the reader. Approaches of this kind will be discussed in Part IV of the book, and especially in section 14.

It might be thought that the graphs-based approach is also suitable for tackling the harder REG problems that will occupy us in the coming chapters, but this is doubtful on closer reflection. Labelled directed graphs as a vehicle for REG have a number of limitations: their focus on logically simple (i.e., atomic) knowledge makes them less suitable for generating logically complex REs, such as "the woman who plays a flute or owns a piano", "the man who loves all dogs", and so on [van Deemter and Krahmer, 2007]. REs of this kind will be discussed in chapter 10, where REG is linked with Knowledge Representation and automated reasoning.

6.6 Appelt's Heirs: Reference as Part of a Wider Problem

The algorithms discussed so far focus squarely on the generation of REs. This narrow focus reflects the methodology, discussed in chapter 4, which isolates reference from other communicative goals and concentrates on determining the semantic content of the REs to be generated. Although this methodology

has allowed researchers to focus on a clear goal, some favour a more holistic approach. Let us summarize the holists' arguments briefly.

First, it is of little use to determine the semantic content of an RE unless this content can be expressed through a grammatical sentence. A doctor who is starting to write "You suffer from an inflammation of the ..." may realize that the organ affected by the inflammation has no commonly known name. Once she realizes this lexical gap, she might re-plan, writing "You suffer from x", if there happens to exist a commonly known name for the condition x. Matthew Stone and Bonnie Webber conclude that REG has to be embedded into a wider algorithm that includes at least lexical choice, and possibly other components of the generation system [Stone and Webber, 1998].

Second, reference and other tasks can help each other. Reference can help with other tasks, for instance, when a speaker says "don't sit at the newly painted table" (an example from Dale and Reiter), where "newly painted" may help to explain *why* the request is made. Conversely, reference is not accomplished by a noun phrase on its own. Consider a domain that contains several hats, only one of which has a rabbit in it; the domain contains other rabbits as well [Stone and Webber, 1998]. Now the utterance "Remove the rabbit from the hat" contains an NP "the hat" that fails to refer if existing algorithms are to be believed; similarly, the NP "the rabbit" appears not to refer uniquely, because there are several rabbits. Yet it is completely clear what hat is to be de-rabbited, and which rabbit will be produced. The authors argue that this is because other parts of the sentence are to be taken into account, and especially the verb "remove": after all, one can only remove a from b if b contains a first; similarly, only something that's in b can be removed from b. The process is similar to the one discussed in section 6.4, in which two parts part of a relational RE (e.g., "the cup on the table") disambiguate each other.

The authors address both problems by deviating from the standard NLG pipeline [Mellish et al., 2006], making reference a more organic part of NLG. The proposed setup, in which Content Determination and Linguistic Realization are interleaved,[9] prevents "ineffable" semantic contents from being generated. Once again, Constraint Satisfaction can be employed to make things work. For example, a syntactic constraint such as "every RE has a noun as its head" can ensure that every RE expresses a TYPE, which is something that other approaches have to enforce artificially.

9 Compare [Horacek, 1997, Krahmer and Theune, 2002, Siddharthan and Copestake, 2004], among others, where these processes are interleaved as well.

The line of research that was outlined very briefly in the above (cf., [Heeman and Hirst, 1995], [O'Donnell et al., 1998], [Koller and Stone, 2007], [Garoufi and Koller, 2014]) follows the tradition started by Douglas Appelt and Amichai Kronfeld, who sought to understand the place of reference within a wider theory of speech acts. It is an apt reminder of the reductive nature of the mainstream program of research on REG, in which important issues can sometimes fall by the wayside. Like the California School, this line of work has the potential to do justice to the place of reference in the wider system of human communication, where reference seldom comes alone, and where reference is not always achieved by an isolated noun phrase. However, this holistic tradition can sometimes resemble older work in less fortunate respects as well: details of data structures and algorithms can sometimes be difficult to glean from published papers, and some theories appear to be driven by isolated examples. This is not always the case. For instance, the ideas discussed above were taken further by Pamela Jordan and Marilyn Walker when they focussed on dialogue [Jordan, 2000a, Jordan, 2000c, Jordan and Walker, 2005].

Jordan hypothesized that REs are affected by multiple communicative goals and tested this hypothesis on the COCONUT corpus [Di Eugenio et al., 2000], which records dialogues between pairs of participants who play a game in which they buy furniture together on a fixed budget. Examples of communicative goals are: committing to the purchase of an item, persuading one's dialogue partner to buy an item, changing a previous commitment, and so on. These goals are defined in terms of recognisable features of a dialogue, enabling annotators to indicate when they apply. For example, a *summarize* context is defined as arising once an agreement has been reached for an action, and ends if the agreement is nullified. Most of Jordan's hypotheses were confirmed [Jordan, 2000b, Jordan, 2002] when the corpus was analysed (see [Jordan and Walker, 2005] for discussion), motivating their inclusion in a new model of reference production, known as the Intentional Influences model.

The model reported in [Jordan, 2000c] starts by accumulating any properties that contribute to the satisfaction of the speaker's communicative goals (e.g., committing to the purchase of an object). Next, the model checks whether the properties accumulated so far identify the referent, adding properties only if they do not, in which case one of the classic REG algorithms is employed to accomplish this (line 13). Unique reference, in other words, is treated like an afterthought. I understand the structure of Jordan's Intentional Influences model ([Jordan, 2000c] section 6.3.2) to be as summarized in Algorithm 10.

Algorithm 10 Jordan's Intentional Influences model

Input: A domain of objects, containing a target referent r and a non-empty set of distractors. A set of mutually known properties of r and a dialogue history.
Output: A contextually appropriate distinguishing description of r if one exists.

```
Find the TYPE of r
if the context is a summarize context then
    let D contain all mutually known properties of r
else if the context is a commit context then
    let D contain all properties of r mentioned in the offering utterance
else if the context is a verify context then
    let D contain all properties of r expressed in the previous turn
else if the context is a persuade context then
    let D contain the properties of r that make of r a better solution
else if the context is a change constraint context then
    let D contain the properties of r that imply the change.
if D does not identify r uniquely then
    apply a classic algorithm to r and add all the resulting properties to D
```

A few years later, Jordan and Walker implemented a version of the model that exploits additional information concerning the dialogue history (e.g., How long ago is it that the referent was last mentioned? Who was the speaker of the previous utterance?). Features from the Intentional Influences and two other models were fed to RIPPER, a Machine Learning device that produces *if-then* rules. A few simplified rules are shown in the following, where C says the RE contains a Colour attribute, O stands for Owner, and Q for quantity [Jordan and Walker, 2005]:

if *goal* = select-chairs and *distance-of-last-state* ≥ 3 and *speaker-of-last-state* = self, then say COQ

if *prev-commit-speaker* = commit and *influence-on-listener* = action-directive and *color-contrast* = no & *speaker-of-last-state* = self, then say C

For example, the first rule says that if the goal of the current utterance is to select chairs and the referent was last mentioned more than 3 utterances ago, in an utterance that had the same speaker as the current one, then colour, owner and quantity (COQ) are expressed. The rules that were learned were shown to perform well, particularly because of features from the Intentional Influences model (such as *prev-commit-speaker and influence-on-listener, goal*).

Jordan and Walker's main claim, that REG algorithms should look beyond the task of identifying the referent, paying attention to a variety of communicative goals, is well taken. In the Epilogue (section 16.2), when we re-visit the Gricean Maxims, we shall argue that theirs is one of the more promising approaches so far to the tricky problem of ensuring that generated REs favour contextually relevant properties.

In this Second Part of the book, we have addressed the classic REG task. In Part III we shall address a more general task. Given is, once again, a finite domain of entities and a target referent. This time around, however, the target referent can be a subset, as well an an element, of the domain. Given is, furthermore, a set of properties that may or may not be logically atomic: they are Logical Forms that may contain various logical operators. If the aim of the algorithm is to model speakers (rather than to optimize utility for hearers) then, in the abstract, the new task can be thought of as follows:

> *The extended* REG *task.* If there exists a *Logical Form* denoting the target, then the REG algorithm needs to (1) find such a Logical Form, making sure that (2) the Logical Form is as similar as possible to the set of properties found in a typical human-authored description. If no distinguishing description of r exists, then the algorithm should say so.

We shall see that the extended REG task, with its greater range of both referents and REs, poses difficult new challenges.

6.7 Summary of the Chapter

This chapter has discussed a number of approaches to the classic REG task that differ from the classic REG algorithms. As well as discussing, in section 6.6, a range of approaches in the tradition of the California School (cf., section 4.2), we have concentrated on the following "alternative" approaches:

- **Probabilistic algorithms.** Recent REG algorithms try to reproduce the probability distribution over the set of REs that are found in a corpus. As a result, they are able to do justice to a wide variety of production behaviours, instead of only the behaviour most frequently observed. Evaluation of these nondeterministic algorithms requires an evaluation method based on computation of the conditional probablity $P(a|b)$, where b is the model expressed by the probabilistic algorithm and a is the data. [Section 6.3]

- **Constraint Satisfaction.** Constraint Satisfaction offers an attractive approach to problems in which different considerations need to be taken into account, because it allows a clear separation between the (declarative) statement of a problem and its (procedural) solution. [Section 6.4]
- **Graphs-based approaches.** Krahmer and his co-workers have used graph-based representations to construct REG algorithms that take account of the *cost* of an RE, where a lower cost can be interpreted as indicative of a more felicitous RE. We have argued that costs allow more fine-grained control than the Preference Orders of chapter 4. These considerations make this approach well suited to the classic REG problem. We shall see in chapter 10 that when harder versions of the REG problem are addressed, richer formalisms are required, which support reasoning with complex information. [Section 6.5]

III THIRD PART: GENERATING A WIDER CLASS OF RES

7 First Extension: Using Proper Names

It is time for us to look beyond the classic REG task, which focusses on one-shot REs, one referent, and one-place predicates. The classic REG task leaves out many kinds of REs, such as proper names, which are the topic of the present chapter. This chapter is shorter and more tentative than most, because little computational research has been done in this area yet. I shall argue, however, that work on the generation of proper names, and complex REs that contain proper names, is urgently needed.

We have seen in section 2.7 that logicians interested in natural language have studied proper names in great depth; linguists and neuroscientists have joined them in later years (e.g., see relevant parts of [van Langendonck, 2007]), and so have computational linguists working on Information Extraction (section 1.2, for example in connection with Named Entity Recognition).

It will be clear from Part II that researchers in REG have paid proper names much less attention. Some have investigated the choice between proper names and other REs [Henschel et al., 2000], [Piwek, 2008], but seldom (with the exception of a simple approach in [Winograd, 1972]) in connection with REG as understood in this book, where Natural Language Generation starts from a Knowledge Base rather than a text. I believe this to be a significant omission, so let us see how proper names might find a proper place within REG. Algorithms play an important role in this book, but the present chapter is a reminder of an important lesson that is well understood in Computing Science: the *representation* of a problem can be the key to its solution, sometimes to such an extent that the algorithms themselves are pushed into the background.[1]

This brief chapter will start by asking why proper names have so far been neglected in REG (section 7.1), after which we shall present a first attempt at letting REG algorithms take proper names into account (section 7.2). A slightly more sophisticated approach, based on the reification of names, is presented in section 7.3. The chapter concludes with a discussion of the challenges posed by proper names, which focusses on the need for empirical results regarding the choice between names and descriptions, and the difficulty of obtaining these results in a sound manner (section 7.4).[2]

1 Representations also tend to take centre stage in approaches based on Constraint Satisfaction, see section 6.4.

2 An initial exploration of the issues that arise when proper names become part of REG was offered in [van Deemter, 2014], section 4.

7.1 Why Have Proper Names Been Neglected in REG?

The usefulness of proper names is not in doubt. Where they are available, names are often the preferred way of referring, particularly when the alternative is a lengthy conjunction of properties (cf., chapter 13). Perhaps the reason why REG researchers have disregarded proper names is that if proper names are allowed, reference is thought to be trivial: to refer to an entity, just generate its proper name! But, on ten seconds' reflection, this is not a defensible position.

For a start, the choice between a proper name and a full description is often pragmatically significant, and so is the choice between different versions of a proper name. References to people come with social implications to do with familiarity, affection, and social hierarchy. Consider a couple with two children. When the son refers to the daughter, he can choose between a proper name and a description. If he wanted to refer to her when addressing his mother, he could say "my sister", "my sibling", "your daughter", "my father's daughter", and so on. Yet, only a proper name would be considered normal. Relation-denoting words like "aunt" complicate matters further, by adding a descriptive element to a name. The expression "Aunt Agatha", for example, would be quite normal to use when referring to someone who is the aunt of the speaker, when addressing herself or another family member (except possibly where that family member is the referent's son or daughter, in which case "your mother" is more appropriate); it would be dis-preferred in most other situations. On top of all these nuances, many languages allow names to be abbreviated (e.g., in English, Christopher becomes Chris; Timothy becomes Tim), with abbreviations carrying subtle implications of familiarity. Clearly, choosing between all these expressive possibilities is far from easy.

Other complexities tend to come to the fore when proper names are used as part of a larger RE, as when we say "Fido the dog", "the poet Burns", or "the River Thames" (examples from [van Langendonck, 2007]). Of particular interest to us are REs in which we refer to an entity *via* some other entity for which we have a proper name. For example, we routinely refer to people as "the CEO of So-and-so" (where So-and-so is a proper name), "the new Principal of So-and-so University", and so on, particularly when the role of the person is more important than his or her name. In some rarer situations, the name may not even be known. Suppose, for example, that, in 1997, you had asked me who was the most influential author of the year 1996, and I responded "the author

of the novel Primary Colors". This would have been a good description of Joe Klein, the political commentator who had anonymously written this book, about a character resembling the then-president Bill Clinton. In my reference to the author, I was unable to use his name but I was able to use the name of his book.

7.2 Incorporating Proper Names into REG

What is needed is for proper names to become part of the REG mechanism, to allow them to be combined with other properties and relations. The simplest way in which this can be done is by assuming that each individual in the Knowledge Base comes not just with a number of descriptive properties but, sometimes, with one or more proper names as well. In other words, a proper name is property. Some individuals may have a proper name and others may not. For simplicity we shall assume that reference by proper name is preferred over reference by properties, so "Joe Klein's dog" is preferred over "The dog of the author of Primary Colors".

This example gets us to another semantic reason for treating proper names seriously: one likes to think of proper names as unambiguous, but in practice, they seldom are. There must be numerous individuals named Joe Klein, for example, which is why I introduced him above as "Joe Klein, the political commentator who ...". One possibility is to regard a proper name as a property that is true of all individuals who have this name. For example,

– (being named) Joe Klein is a property of all individuals named Joe Klein
– (being named) Joe is a property of all those individuals named Joe
– (being named) Klein is a property of all those individuals named Klein

The idea that proper names are properties is far from new and not without its modern defenders. [Burge, 1973], for example, invoked sentences like

– There are relatively few Alfreds in Princeton.
– An Alfred joined the club today.

[Larson and Segal, 1995] added further arguments in support of the idea that names are properties. Elbourne, who reviews and augments these arguments, observes that we can talk about "*that* Alfred", and that languages such as classical Greek and modern German allow the combination of a definite determiner

with a proper name ("Der Alfred") and defends Burge's perspective against various objections [Elbourne, 2005].

Of course, in realistic settings where domains may be large, REG should come with a mechanism for the management of salience. As we have seen in sections 1.6 and 4.9, an RE like "the dog" does not really denote the only dog in the universe – or else we could never use this expression felicitously – but only the *most salient* dog. Thus, a REG algorithm has reached its aim of identifying a referent r if it has produced a description $\mathcal{D}$ such that the most salient element of $\mathcal{D}$'s denotation $[\![D]\!]$ is r. We assume here that the same idea applies to proper names, thus, "Joe Klein" will individuate an entity if it is the most salient entity in the Knowledge Base whose name is Joe Klein. Similarly, "the political commentator Joe Klein" will be considered a legitimate reference to r if r is the most salient political commentator whose name is Joe Klein. To make this work for *parts* of names as well as full proper names, the easiest approach is to associate a *set* of names with each individual, for example:

OCCUPATION: political commentator
NATIONALITY: American
NAMES: {Mr Joe Klein, Joe Klein, Joe, Klein}

Because longer versions of a person's name will be applicable to only some of the individuals to whom a shorter version is applicable, the different values of the NAMES attribute will tend to subsume each other: all people who are called Mr Joe Klein are also called Joe Klein, and all of these are called both Joe and Klein. These properties could consequently be dealt with using the mechanism for subsumption in the Incremental Algorithm, which can account for the fact that all dogs are canines, all canines are mammals, and so on (see the function `FindBestValue` in section 4.6).

We shall not discuss which types of individuals tend to have commonly known proper names (people, cities, and companies come to mind) and which do not (e.g., tables, trees, body parts, atomic particles). Likewise, we will say little about the choice between different versions of a proper name (e.g., given name, surname or both), a difficult issue that is handled differently in different languages and cultures, and which can sometimes depend on the length and frequency of the names involved. The social implications of titles and honorifics complicate these questions even further; for example, to say "Mr Klein" is more than simply to refer to the most salient male individual named "Klein", but also to do this politely.

To sum up the core of approach outlined so far:

- Each individual is associated with an attribute NAMES.
- For a given individual, the set of values of NAMES can be empty (no name is available), singleton (one name), or neither (several names).
- A subsumption relation can be defined among these values.
- Different individuals can share some or all of their names (cf., section 1.2, where entity resolution is discussed).
- REG will treat the NAMES attribute in the same way as other attributes.

It appears that names are often the "canonical" way of referring to an entity. If this is true, then standard mechanisms could be invoked to favour names at the expense of other properties, including Dale and Reiter's Preference Order, or even (given that names are often short) a preference for brevity. This can be likened to the idea that numbers can be written in canonical form by writing them in decimal form: $\sqrt{16}$ denotes a *bona fide* number, but only 4 is the canonical form of that number. But just as there can be reasons for writing $\sqrt{16}$ (and certainly $\sqrt{2}$), there can be reasons for not using proper names. If you know the name of the woman involved, this does not mean that "Please contact the Director of the Customs and Tax Department" is better worded as "Please contact Mr X" (where X is her name). In referring to sets (chapter 8), it is even less clear that proper names are always preferred, because listing proper names does not necessarily make for a short description (compare "the citizens of China" with a listing of all the elements of this set). Once again, there are many open questions about these matters.

To see how things could work in practice, suppose the facts on the ground are as follows. For simplicity, each individual has exactly one name:

TYPE: woman $\{w_1, w_2, w_3\}$, man $\{m1\}$, dog $\{d_1, d_2\}$
NAMES: mary $\{w_1\}$, shona $\{w_2, w_3\}$, rover $\{d_1\}$, max $\{m_1, d_2\}$
ACTION: feed $\{(w_1, d_1), (w_2, d_2), (w_2, d_1)\}$
AFFECTION: love $\{(w_1, d_1), (w_3, d_1)\}$

Then our proposal suggests the following referential possibilities:

d_1: "Rover"
d_2: "The dog called Max" (Just "Max" could refer to m_1.)
w_3: "Shona, who loves a dog" ("Shona" could be w_2,"loves a dog" could be w_1.)

7.3 Reifying Properties

The representational scheme that we use here is fairly flexible; for example, it would allow a suitable generator to quantify over people using their names, as when we say "There are two Shona's in this domain". With this representation scheme in place, classic REG algorithms can be applied without any modifications. However, as Graeme Ritchie (p.c.) has pointed out, the representation scheme does not allow proper names to have properties (e.g., "is a posh name", "has 5 characters" ,"is common in Scotland"). If names are *reified*, then this becomes possible; what's more, proper names themselves could be referred to (e.g., "the name his friends call him"). This idea, which could be generalized to other properties ("is a nice colour", "is the colour of grass", etc.), can be worked out in different ways, but let me sketch one possibility.

The idea is to treat a name as just another object linked (on the one hand) to the things it names and (on the other hand) to the ways in which it manifests itself in spelling (pronunciation, etc.). One name, n_2 for example, may name both a man and a dog, and it is written as "Max". This would lead to the Knowledge Base above being expanded as follows:

Type: woman $\{w_1, w_2, w_3\}$, man $\{m_1\}$, dog $\{d_1, d_2\}$, name $\{n_1, n_2, n_3, n_4\}$
Action: feed $\{(w_1, d_1), (w_2, d_2), (w_2, d_1)\}$
Affection: love $\{(w_1, d_1), (w_3, d_1)\}$
Naming: name $\{(d_1, n_1), (d_2, n_2), (w_1, n_3), (w_2, n_4), (w_3, n_4), (m_1, n_2)\}$
Spelling: written $\{(n_1, Rover), (n_2, Max), (n_3, Mary), (n_4, Shona)\}$

The revised Knowledge Base can do everything the previous one can, and additionally it can be used to generate REs such as "The name shared by a man and a dog" (to refer to the name written "Max"). If n_4 is also specified to be Scottish, it can generate REs such as "Those women who have a Scottish name" as well. The only drawback of this approach, in which names are objects rather than values of an attribute, is that subsumption of values can no longer be used to express the relationship between, for example, "Klein" and "Joe Klein". Future accounts would have to address the fact that not only individuals can have names, but sets of individuals as well (cf., chapter 8). These complexities become particularly clear if we move away from people to geographical areas, for example, where there are names for individuals at many levels: Rosemount is an area of Aberdeen, which is the capital of Aberdeenshire, Scotland, for instance, and this hierarchy opens up a wealth of referential possibilities.

7.4 Challenges for REG Posed by Proper Names

Proper names would make an interesting case study into the relation between philosophical theory and computational linguistics practice. The analyses above owe little to work in the philosophy of language, where some of the best minds have studied the logic of proper names, puzzling over the ways in which their behaviour differs from that of definite descriptions, for instance in epistemic contexts (see section 2.7).

Epistemic contexts have so far been ignored by computational models, but it is not obvious that our account fares worse in such contexts than philosophers' accounts. A sentence like "Aristotle could have had a different name", for example, can be dealt with by the two accounts above if different possible worlds are associated with different Knowledge Bases: the sentence could be analysed as meaning that the object named "Aristotle" in the actual world has a different name in some other worlds.

Admittedly, if Kripke was right, then an example like "Aristotle might not have been Aristotle" is harder to deal with in the models outlined above. But was Kripke right to think this sentence is necessarily false? Existing work on proper names (section 2.7) has not yet reached the stage where there is unanimity about the facts, let alone their analysis, particularly in modal and epistemic contexts (see e.g., section 2.5). Until good theories of such contexts are available, computational models could justifiably focus on the issue of expressive choice, which lies at the heart of Natural Language Generation (recall section 1.3). This work could focus on questions of style, formality, and politeness, for instance, where many open questions lend themselves to experimental investigation [van Langendonck, 2007].

This chapter has only scratched the surface of a large problem: we have only discussed names of individual entities, but sets (chapter 8) can have names as well (e.g., "The Rolling Stones"), and so can geographical areas (chapter 14), for instance. A proper treatment of proper names should permit the use of names in all such cases, perhaps particularly as parts of a larger RE, and in combination with vague expressions (e.g., "the mountains just west of Barcelona").

Future experiments on the generation of REs that involve proper names could address the empirical questions in this area. To find out when proper names are used, one could perform an experiment along the lines of the TUNA's people corpus (section 5.3), where speakers were invited to refer to stimuli on a

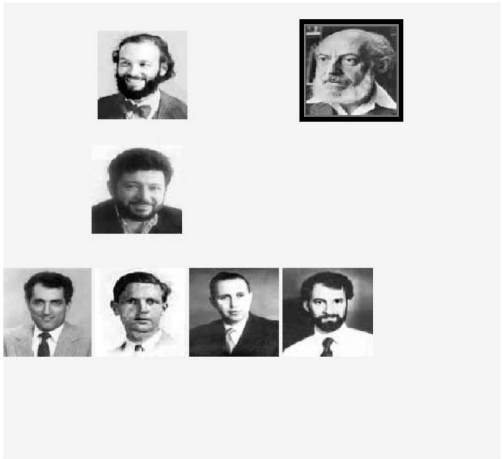


Figure 7.1
A trial in the "people" part of the TUNA *experiment (rendered here in black and white).*

screen. The person highlighted in the top right of Figure 7.1, for example, could be referred to as "the gentleman with the white beard" or as "Professor Samuel Eilenberg" (a one-time mathematician at Columbia University). It would, however, have been very difficult to ensure that proper names compete with other properties on a level playing field, neither favouring nor disfavouring proper names.

To appreciate the difficulty, note that the participants in the original TUNA experiment could have used proper names already, because the people in the pictures depict actual persons (famous mathematicians, in fact). This never happened, however, because our participants were not familiar with these names. To make them familiar with the names of the referents, we could have trained our subjects before the experiment. But what level of familiarity would have been appropriate? If the training had been very extensive (or if we had added these names as captions to the images, for example, as in the map task corpus of [Anderson et al., 1991] [Bard, 2007]), then this would have caused such a strong association between the name and the referent, in the minds of the participants, that it would have biassed their utterances hugely towards using names, at the expense of descriptions. The root of the problem is that proper names are conventional in a way that descriptive expressions (e.g., "with a white beard") are not, and this fundamental difference makes it difficult to

compare the two types of REs fairly. This is possibly the hardest challenge posed by proper names.

One final consideration is related to the issue of conventionality. In our initial discussion of Information Sharing (section 1.7), we saw that the normal direction of sharing can sometimes be reversed, as in Barwise and Perry's example in which a speaker says "My wife is coming in just now", thereby informing the hearer that the person who is walking into the room *is* his wife. Likewise, if I use a proper name n to refer to the referent r, I tell the hearer that I consider the fact that n is a name of r to be in our common ground. This is unremarkable in many cases, but especially when n is a nickname it can be crucial: by saying to you that "The Gunners" have just won a football match (this is a nickname for the North-London football club Arsenal), I'm telling you that this nickname is in our common ground. Effectively, I am thereby saying that we both belong to a social circle of people who are "in the know" about the club. Computational models are not yet able to formalize under what circumstances a nickname is an appropriate way to refer, and how it affects the shared (social) information of the interlocutors.

7.5 Summary of the Chapter

This chapter argues that REG algorithms should take proper names into account and offers a preliminary discussion of what might be the best way to do this. We discuss two closely related approaches, both of which are variants of the well-known theory that proper names are disguised definite descriptions (cf., section 2.7).

- The first proposal treats a proper name as a property of the referent. It appears to account for many of the basic facts about proper names. In particular, a referent can have several proper names (some of which can be more specific than others); a proper name can be ambiguous; a name can be combined with other properties of the referent. [Section 7.2]
- The second proposal treats a proper name as an object to which a referent can stand in the naming relation, and which can be written and pronounced in various ways. This approach accounts for a number of additional facts, such as the fact that a name can have properties and that a name can be referred to. [Section 7.3]

- Neither of these proposals has much to say about the behaviour of proper names in modal or epistemic contexts, which have figured strongly in theoretical discussions of proper names (section 2.7). We have argued, however, that this is to be expected given the state of our understanding of proper names. [Section 7.4]
- The hardest problems in the area of this chapter include choosing between proper names and other REs, choosing between different versions of the same name, and figuring out how names are best combined with epithets and other properties. These questions could be studied using the type of elicitation experiments discussed in Part II. However, because names are purely *conventional* devices, it is difficult to design an elicitation experiment that compares names with other REs on a level playing field. [Section 7.4]
- Even though the issues discussed (briefly) in this chapter are particularly relevant for proper names, they have some relevance for other types of REs as well. This is true for reification of properties (section 7.3), for the issue of politeness (after all, not all properties of a person are equally suitable for use in a politely used reference to the person), and even for the question of conventionality (section 7.4). The study of proper names, and REs containing them, can be a useful testing ground for future REG models that take these difficult issues into account.

8 Second Extension: Referring to Sets

Classic REG algorithms do not only bypass proper names: they simplify in other ways as well. For example, a referring expression (RE) produced by such an algorithm can only refer to one single entity, never to a larger set. Moreover, the classic algorithms are restricted to using logical conjunction (i.e., set intersection) as the only mechanism through which properties may be combined. This chapter will explore the questions that come up when these two related restrictions are abandoned. Reference to sets will prove to be a surprisingly rich area, full of little-explored questions and open problems, some of which will be dealt with in later chapters.[1]

The plan for this chapter is as follows. We start discussing how some references to sets can be produced by a simple variant of the classic algorithms (section 8.1). Then we discuss some limitations of this simple approach (section 8.2), followed by two ways in which a wider range of references to sets can be generated (sections 8.3 and 8.4). Then we move on to experimental studies, focussing on two problems that arise from the fact that references to sets are often *conjunctive*, being of the form "the so-and-so and the so-and-so". The first problem is that the two parts of such a conjunctive reference can sometimes be ill matched, because each conjunct applies a different conceptual perspective (section 8.7). The second problem is that conjunctive references can give rise to syntactic ambiguities, as when we say "the old men and women", where it is unclear whether this could include any young women (section 8.8).

Because sets are at the centre of this enterprize, the use of set-theory notation is difficult to avoid. Accordingly, the presentation in this chapter, and in the later chapters of Part III, will be more formal than elsewhere in this book.

Readers who are particularly interested in empirical issues relating to reference to sets should be able to jump to sections 8.7 and 8.8 without difficulty.

8.1 Purely Conjunctive References to Sets

Suppose the information shared by a speaker and hearer is as captured by the following Knowledge Base, whose domain is a set of dogs, and whose only attributes are TYPE and COLOUR:

1 This chapter integrates ideas from [van Deemter and Halldórsson, 2001], [van Deemter, 2002], [Gatt and van Deemter, 2007], and [Khan et al., 2012] with more recent work, using the monotonic approach to REG outlined in chapter 4.

TYPE: Dog ($\{a, b, c, d, e\}$), Poodle ($\{a, b\}$)
COLOUR: Black ($\{a, c\}$), White ($\{b, e\}$)

What would be more natural than to refer to $\{a, b\}$, which is a set, as "the poodles", or to $\{b, e\}$ as "the white dogs"? Reference to sets is often neglected, throughout the Cognitive Sciences. Yet generating descriptive REs is even more crucial if the target is a set than if it is a single object: even if the objects in the set have proper names, the set as a whole is likely to lack a name. (Each of the occupants of the room next to mine has a proper name but, as a group, they do not.) Moreover, if the set is large, then enumerating its elements is cumbersome. In this section, we sketch extensions of REG that produce distinguishing descriptions of sets. We shall see that reference to a set is a more complicated affair than reference to an individual object.

In the examples considered so far, we are lucky, because the elements of the target set happen to have something in common that they share with no other entity. In this case, we can generate referring expressions in the same way as before, searching for atomic properties $P_1, ..., P_n$ whose intersection equals a given target set. For example, we could aim for the shortest conjunction that singles out the target set. Or, we can use an incremental strategy, as in the algorithm below, which we call $\text{IA}_{\text{Plural}}$. As often before, we disregard special provisions for head nouns (e.g., *via* the TYPE attribute). Note, however, that a head noun must be selected that suits every element of the target set.

The new algorithm generalizes the original Incremental Algorithm: S takes the place of the target referent r. The process of expanding L and contracting C continues until all domain objects that are not elements of S have been removed. $\text{IA}_{\text{Plural}}$ refers to a set by scrutinizing its elements. Because S may be a singleton, $\text{IA}_{\text{Plural}}$ subsumes IA. As before, we assume a nonempty set of distractors. In Algorithm 11, we state the algorithm informally, using the monotonic structure familiar from our discussion of the Greedy and Incremental Algorithms for reference to a single object (chapter 4).

This simple algorithm was tested as part of the TUNA experiment of chapter 5, with results that were nearly as good as those for reference to individual entities [van Deemter et al., 2012b]. (For the furniture corpus, in situations where a nondisjunctive RE is able to single out the referent set, the Dice score for the best performing Incremental Algorithm was 0.797; for the people corpus this figure was 0.819.) An example domain is shown in Figure 8.1.

Algorithm 11 IA$_{\text{Plural}}$: the Incremental Algorithm for referring to sets

Input: A domain of objects, containing a non-empty target set S consisting of elements of the domain; also containing a non-empty set M of distractors (i.e., domain elements that are not elements of S). A set $\mathcal{P}$ of properties true of each element of S. A linear Preference Order defined on $\mathcal{P}$.
Output: A distinguishing description $\mathcal{D}$ of r if one exists.

1: Start out with an empty $\mathcal{D}$
2: **while** Not all distractors have been ruled out and $\mathcal{P} \neq$ empty **do**
3: Select a new property P from $\mathcal{P}$, *choosing the most preferred one*
4: **if** P is false of some distractors **then**
5: Add P to $\mathcal{D}$
6: Remove P from $\mathcal{P}$
7: Remove from M all distractors ruled out by P

(a) People trial

Figure 8.1
The TUNA *elicitation experiment (people): reference to sets.*

Collective properties. As an aside, note that this approach does not work for collective properties, such as "being of the same age", which pertain to *sets* of objects (e.g., [Scha and Stallard, 1988], [Lønning, 1997]). A REG algorithm that exploits collective properties was first proposed in [Stone, 2000], using a constraint-based approach. The monotonic algorithm $\text{IA}_{\text{Plural}}$ can be modified to do the same.[2] This time around, M is a set of sets and $\mathcal{P}$ is a set of collective properties true of the target referent set.

Algorithm 12 IA_{Col}: the Incremental Algorithm, referring to sets and allowing collective properties

Input: A domain of sets, containing a target set S and a non-empty set M of distractors (i.e., elements of the domain that are not identical to S). A set $\mathcal{P}$ of properties true of S itself. A linear Preference Order defined on $\mathcal{P}$.

Output: A purely conjunctive distinguishing description $\mathcal{D}$ of S if one exists.

Steps: Identical to the steps of $\text{IA}_{\text{Plural}}$ (Algorithm 11).

IA_{Col} selects properties of *sets*, removing from the set of all potential referents (i.e., the set of all subsets of the domain) all those sets for which the property is false. For example, selection of "being of the same age" removes all sets whose elements are not of the same age as each other; selection of "football team" removes all sets that do not make a (complete) football team. The algorithm generates Logical Forms for descriptions such as "football teams whose members are of the same age".

As it stands, IA_{Col} applies to collective properties only. However, the algorithm can take distributive properties in its stride if these are upgraded to the level of sets (cf., [Kamp and Reyle, 1993], p. 338): let P be a distributive property, then we define the collective property $d(P)$ to be true of a set if and only if P is true of all elements of the set. For instance, suppose $S = \{a, b, c\}$ is a team of a 3-player sport and S is the only such team all of whose members have the flu (i.e., $Flu(a), Flu(b), Flu(c)$). This makes the property $d(Flu)$ true of S. The algorithm above can single out S by combining the collective property of being a team with the distributive property $d(Flu)$. The target set S is identified as the team whose members have the flu. (End of aside.)

2 These adaptations would cause the worst-case run time of the algorithm to become exponential, because testing whether all distractors have been ruled out involves inspecting all subsets of the domain, of which there can be up to 2^{n_d}, where n_d is the cardinality of the domain.

Figure 8.2
The need for negation (1): "I bought the two dogs that are not poodles".

8.2 Negation and Disjunction

From here on, we focus on the case in which all properties ascribed to a referent set are distributive. Before we proceed, one terminological issue should be clarified: Perhaps confusingly, disjunction ($\vee$) (or its equivalent $\cup$ in set theory, if you prefer) may be realized in language as conjunction ("and"). For example, the set $\{x \mid x \in A \vee x \in B\}$ can be referred to as "the As *and* the Bs". Reversals of this kind occur in other linguistic constructs as well. For example, "You may do A *or* B" equals "You may do A *and* you may do B", a phenomenon known as Free Choice Permission (e.g., [Asher and Bonevac, 2005]).

Now that we are able to generate references to sets, let us move away from purely conjunctive descriptions to full Boolean combinations of properties. First look at Figure 8.2, where a speaker says "I bought the two dogs that are not poodles", which contains a negation; a better description seems hard to find. A simpler example, where negation is unavoidable and where the ref-

erent is one single individual, can be constructed if we consider the following Knowledge Base, a variant of the one in section 8.1:

TYPE: Dog ($\{a, b, c, d, e\}$), Poodle ($\{a, b\}$)
COLOUR: Black ($\{a, b, c\}$), White ($\{d, e\}$)

Classic REG algorithms, such as the IA, do not allow us to individuate c. Yet c is unique: it is the only black dog that is *not* a poodle: $\{c\} = Black \cap \overline{Poodle}$. Let's reflect on this situation: here is an element which an English speaker would have no difficulty singling out by means of an NP; yet none of the algorithms discussed so far is able to produce a distinguishing description of c.

One might be tempted to see this as a defect of the Knowledge Base: if for each property P in the Knowledge Base there exists a property coextensive with the complement of P, then the problem disappears. (If c, d, and e are all alsatians, then c is "the black alsatian".) This does not make sense, however, because natural language does not have a word for every complement.

Moreover, gaps in our shared knowledge are only too real: it may be unclear what type of dog c is, for example, because c is a cross between types. Negation can help in such cases. When relations are considered, the advantages of negation become even more pronounced, because negated relations are often not lexicalized. A dog that is not a poodle (as in Figure 8.2) is presumably some other kind of dog, for which we may have a word if we are lucky; but how about a man who is not wearing a sombrero: do we have a word for that? (See Figure 8.3.) Chapter 10 will tackle situations of this kind.

There is a problem, however. Recall that the Knowledge Base is meant to represent information that is in *common ground*. In other words, the reason why c's type is not listed as "poodle" might not be that c is not a poodle, but merely that c's being a poodle is not in common ground. It might even be that c is a poodle, and that both the speaker and the hearer know it, but the hearer fails to know that the speaker knows it. In this case, "the black dog that's not a poodle" will misfire, because to the hearer, there is no such dog. Clark and Marshall's reasoning about two people aiming to watch a film together (section 3.1) showed that there is no simple way around this problem, and that reference can fail. The simplifying assumption that we shall make here is that *all* the information that is relevant to the appropriateness of an RE is in common ground. Focussing on the example above, we assume that it is in the common ground what the elements of the domain are, what their types are, and what their colours are. More generally, we do not only assume that all the propositions listed explicitly in the Knowledge Base are in common ground, but also

Figure 8.3
The need for negation (2): "Arrest the man who is not wearing a sombrero".

the *negated* propositions that follow from the Closed World Assumption (section 1.1). These assumptions can fail, but they tend to be fulfilled in situations where speaker and hearer are co-present.

Before we discuss how REG can deal with these issues, we need to address another limitation of classic REG algorithms, which is their inability to express disjunction (i.e., set union). Consider the example domain above once again, focussing on the set $\{a, b, d, e\}$. It is easily described in set-theoretic terms, for instance, $poodle \cup white$. Moreover, the set *is* referable in English, as when we say "The poodles and the white dogs". Just like negated properties, one can view disjunctions between properties as being implicit in simple Knowledge Bases of the kind on which we are focussing in this chapter. Unlike negation, disjunction does not lead to conceptual problems related to common ground.

In the next two sections, we will investigate how REG algorithms can produce REs that use negation and/or disjunction. But before we talk about algorithms, it will be useful to ask how one can determine whether unique identification of an entity is possible in a given situation. Before negation and disjunction made their appearance this was an easy question but, as the logical apparatus at the disposal of REG grows, this question becomes harder to answer. This same question – which will be discussed more fully in sections 10.3 and 11 – will turn out to give us a useful perspective on reference, ultimately suggesting alternative REG algorithms.

8.3 Satellite Sets and Their Use in REG

Sometimes it is impossible to identify a given individual uniquely; every Logical Form that describes it describes one or more other individuals as well. Following [van Deemter and Halldórsson, 2001], we call these unwanted individuals *satellites* of the referent. Let's now call the *satellite set* of an object r the set of objects from which r cannot be distinguished (this time including r itself). We shall see that the idea of a satellite set can inform a primitive algorithm for referring to sets. Later in this chapter we shall be concerned with other algorithms, but it will be enlightening to see what can be done with the simple idea of a satellite set. Satellite sets will be reused in chapter 10 (where they will be more rigorously defined), so it will be useful to keep track of them during the course of our narrative.

First, consider the classic REG task of referring to an individual. There is a domain $\mathcal{D}$ with a target referent r in it and a number of distractors. Furthermore, there is a set $\mathcal{P}$ of atomic properties. Consider the intersection of the extensions of all properties that hold true of a referent r:

$$\bigcap\{[\![P]\!] : P \in \mathcal{P} \wedge P(r)\}$$

In section 4.3 we encountered the Total Reference algorithm (Algorithm 2), which essentially follows the structure of this formula. The algorithm calculates, for each of the properties P in $\mathcal{P}$, its extension $[\![P]\!]$, assembling these properties in a list; it remove from the list the properties P for which $r \notin [\![P]\!]$, resulting in a new list. The algorithm then computes the intersections of the extensions of all the properties in this list. If the resulting intersection equals the singleton $\{r\}$, then the tentative Logical Form is an RE that identifies r; if it is not, then an intersective RE does not exist.

Three factors need to be distinguished: the Logical Form, its extension, and the algorithm inspired by it. We call the Logical Form the *satellite set constructor*, and its extension (which is also the extension of the Logical Form) the *satellite set* of r. Let's return to the shared Knowledge Base of section 8.1, which we repeat here:

TYPE: Dog ($\{a, b, c, d, e\}$), Poodle ($\{a, b\}$)
COLOUR: Black ($\{a, c\}$), White ($\{b, e\}$)

We carry out the above construction for each element of the domain. To simplify presentation, we will omit brackets, writing dog instead of $[\![dog]\!]$, for example; analogously we shall sometimes talk about, for instance, the intersection of properties where we mean the intersection of their extensions.

$SatelliteSet(a) = dog \cap poodle \cap black = \{a\}$
$SatelliteSet(b) = dog \cap poodle \cap white = \{b\}$
$SatelliteSet(c) = dog \cap black = \{a, c\}$
$SatelliteSet(d) = dog = \{a, b, c, d, e\}$
$SatelliteSet(e) = dog \cap white = \{b, e\}$

These simple calculations show that only a and b can be individuated. If your aim is to identify c, and the four properties listed above are the only ones available, then you should give up, because no algorithm will be able to do it. If your aim is to identify the *set* $\{a, c\}$, however, then the intersection of $black$ and dog will do it. (The property dog is logically superfluous in this case, but it is possible to systematically omit superfluous properties in order to make the generated REs less verbose.)

This is not the end of the story, however. This can be seen in the Knowledge Base above, where it is impossible to refer to $\{a, b, d, e\}$ with just conjunction and complementation (i.e., negation): if we add full set complementation, expressed here by means of a horizontal bar, we can express $\overline{black \cap \overline{poodle}}$, or set union, using the Logical Form $poodle \cup \overline{black}$ ("the poodles and the ones that are not black"). One approach to the generation of REs of this kind uses a generalization of the satellite set constructor above. The idea is to intersect all the literals (i.e., atomic properties and their complements) that hold true of each of the n element of the set, and to form the union of these n intersections [van Deemter and Halldórsson, 2001]. This idea can be formalized elegantly using set operators. S is the target set which the algorithm seeks to refer to:

Logical Form = (the formula) $\bigcup_{d \in S} \bigcap_{A \in S_d} [\![A]\!]$,
where $S_d = \{A : A \in \mathcal{P}_{plusneg} \wedge d \in [\![A]\!]\}$,
where $\mathcal{P}_{plusneg} = \mathcal{P} \cup \{\ \overline{P_i} : P_i \in \mathcal{P}\ \}$.

The algorithm inspired by this idea (Algorithm 13 below) starts adding to $\mathcal{P}$ the properties whose extensions are the complements of those in $\mathcal{P}$. The resulting set is called $\mathcal{P}_{plusneg}$. For each domain element d of S, the algorithm finds those properties in $\mathcal{P}_{plusneg}$ that are true of d, and the intersection of the extensions of these properties is formed and dubbed $Sat(d)$. A Logical Form for a referent set S is constructed by forming the union of all the $Sat(d)$, for each d in S. A Logical Form is successful if it evaluates to the target set S; otherwise the algorithm returns *Fail*. If *Fail* is returned, then no Boolean description of S is possible. An example of the working of the algorithm will be given presently.

Algorithm 13 Description by Satellite Sets (DBS)

Input: A domain of objects, containing a non-empty target set S consisting of elements of the domain; also including a non-empty set of distractors (i.e., domain elements that are not elements of S). A set $\mathcal{P}$ of properties true of each element of S. The notation S_d is explained in the text.
Output: A description $\mathcal{D}$ that denotes the set S, if such a description exists.

1: **for** each element d of S **do**
2: construct S_d (which is a set of sets of domain elements)
3: $Sat_d :=$ the intersection of all the sets in S_d
4: $\mathcal{D} :=$ the union of all these Sat_d (i.e., where d is any element of S)
5: **if** $[\![\mathcal{D}]\!] = S$ **then**
6: **return** $\mathcal{D}$

The Logical Forms produced by DBS are lengthy. To identify the target set $S = \{c, d, e\}$, for example, the property $\overline{poodle}$ would have sufficed. The algorithm, however, starts from the three elements of S, performing line 2 of the Algorithm three times:

$S_c = \{dog, black, \overline{poodle}, \overline{white}\}$.
$S_d = \{dog, white, \overline{poodle}, \overline{black}\}$.
$S_e = \{dog, white, \overline{poodle}, \overline{black}\}$.

Consequently, a lengthy disjunction $\mathcal{D}$ is generated:

$(dog \cap black \cap \overline{poodle} \cap \overline{white}) \cup$
$(dog \cap white \cap \overline{poodle} \cap \overline{black}) \cup$
$(dog \cap white \cap \overline{poodle} \cap \overline{black})$.

DBS is computationally cheap: it has a worst-case running time of $O(n.p)$, where n is the number of objects in S and p is the number of atomic properties. Rather than searching among many possible unions of sets, a target $S = \{s_1, .., s_n\}$ is described as the union of n Satellite sets, each of which equals the intersection of those (at most p) sets in $\mathcal{P}_{plusneg}$ that contain s_i. Descriptions can make use of the Satellite sets computed for earlier descriptions. Satellites sets can even be calculated off-line, for all the elements in the domain, increasing computational efficiency even further.

The algorithm's profligacy makes this theorem straightforward to prove:

Theorem 2 Boolean Completeness: For any set S, S can be denoted by a Boolean combination of properties in $\mathcal{P}_{plusneg}$ if and only if $\bigcup_{d \in S}(\text{SatelliteSet}(d))$ equals S.
Proof: The implication from right to left is obvious. For the converse implication, suppose that $S \neq \bigcup_{d \in S}(\text{SatelliteSet}(d))$. Then for some $e \in S$, SatelliteSet(e) contains an element e' that is not in S. But $e' \in$ SatelliteSet(e) implies that every property in $\mathcal{P}_{plusneg}$ that holds true of e must also hold true of e'. It follows that S, which contains e but not e', cannot be obtained by a combination of Boolean operations on the sets in $\mathcal{P}_{plusneg}$. □

The completeness of DBS follows directly; the limitation to finite sets stems from the fact that DBS addresses the elements $d \in S$ one by one:

Theorem 3 Completeness of DBS: Assume there are finitely many properties. Then if an individual or a finite set can be individuated by any Boolean combination of properties defined on the elements of the domain, then DBS will find such a combination. □

Algorithms based on satellite sets are fast and elegant. It is time, however, to turn to another way of generating references to sets, which is closer in spirit to the monotonic REG algorithms discussed in earlier chapters.

8.4 Generating Boolean Logical Forms Incrementally

The DBS algorithm generates unwieldy descriptions. Let's see how Logical Forms that refer to sets can be generated incrementally, producing output that is a bit more in line with human language production. We start with the computational mechanism, postponing discussion of empirical issues.

First, we add negations to the list of atomic properties taken into account by an Incremental Algorithm. Then we let $\text{IA}_{\text{Plural}}$ run a number of times: in Phase 1, the algorithm is performed using all positive and negative literals; if this ends before all distractors have been ruled out, Phase 2 removes further distractors

from C by making use of negations of intersections of two literals, and so on, until all distractors have been ruled out or all combinations have been tried. Negation of an intersection comes down to set union, because of De Morgan's Law: $\overline{P_1 \cap ... \cap P_n} = \overline{P_1} \cup ... \cup \overline{P_n}$. Thus, Phase 2 deals with disjunctions of length 2, Phase 3 with disjunctions of length 3, and so on. Optimizations may be applied. For instance, a Logical Form of the form $(P \cup Q) \cap (P \cup R)$ can be simplified to $(P \cup (Q \cap R))$ using standard algorithms (e.g., McCluskey 1965). More will be said about these optimizations in section 8.5.

As usual, a schematic presentation (Algorithm 14) will be useful, in which $P_{+/-}$ stands for any literal, that is, any atomic property or its negation. (Different occurrences of $P_{+/-}$ denote potentially different literals.) The *length* of a property will equal the number of literals occurring in it. We will say that a IA$_{\text{Plural}}$ phase *uses* a set of properties X if it loops through the properties in X (i.e., X takes the place of $\mathcal{P}$ in the original IA$_{\text{Plural}}$). Recall that in IA$_{\text{Plural}}$, $\mathcal{D}$ is the description under construction (which grows during execution of the algorithm), and M is the set of distractors (which shrinks during execution).

Algorithm 14 Generating Boolean descriptions incrementally

Input: A domain of objects, containing a non-empty target set S consisting of elements of the domain; also including a non-empty set M of distractors (i.e., domain elements that are not elements of S). A set $\mathcal{P}$ containing n properties true of each element of S.
Output: A description $\mathcal{D}$ that denotes S, if such a description exists.

1: **Phase 1**: Perform IA$_{\text{Plural}}$ using *all properties of the form* $P_{+/-}$
(and updating M and $\mathcal{D}$ in the process)
2: **if** this is successful **then**
3: **return** $\mathcal{D}$
4: **else**
5: Go to Phase 2
6: **Phase 2**: Adding to the values of $\mathcal{D}$ and M coming out of Phase (1),
Perform IA$_{\text{Plural}}$ using *all properties of the form* $P_{+/-} \cup P_{+/-}$
(and updating M and $\mathcal{D}$ in the process)
7: **if** this is successful **then**
8: **return** $\mathcal{D}$
9: **else**
10: Go to Phase 3. *(Etcetera*, up to Phase n.)

We require without loss of generality that no property, considered at any phase, may have different occurrences of the same atom. (For example, it is useless to consider $\overline{P_1} \cup \overline{P_2} \cup P_1$, which is true of every element in the domain, or the property $\overline{P_1} \cup \overline{P_2} \cup \overline{P_1}$, which is equivalent to the earlier-considered

property $\overline{P_1} \cup \overline{P_2}$.) Because, at Phase n, there is room for properties of length n, the maximal number of phases equals the total number of atomic properties.

To see how this works, consider our example of section 8.2 again:

TYPE: Dog ($\{a, b, c, d, e\}$), Poodle ($\{a, b\}$)
COLOUR: Black ($\{a, b, c\}$), White ($\{d, e\}$)

Suppose the Preference Order of atomic properties corresponds with the order in which they are listed, where the same order extends to their negations, which are less preferred. Abbreviating $B = black$, $D = dog$, $P = poodle$, and $W = white$, this yields the Preference Order $\langle B, D, P, W, \overline{B}, \overline{D}, \overline{P}, \overline{W} \rangle$. Now if $S = \{c, d, e\}$ or $S = \{c\}$ are referred to, a Logical Form is found during Phase 1: $\overline{P}$ in the first case, $B \cap \overline{P}$ in the second. The situation gets more interesting if $S = \{a, b, d, e\}$, which triggers Phase 2. Suppose the properties relevant for Phase 2 are ordered as follows:

$\langle\, B \cup D, B \cup P, B \cup W, D \cup P, D \cup W, P \cup W, B \cup \overline{D}, B \cup \overline{P}, B \cup \overline{W},$
$D \cup \overline{B}, D \cup \overline{P}, D \cup \overline{W}, P \cup \overline{B}, P \cup \overline{D}, P \cup \overline{W}, W \cup \overline{B}, W \cup \overline{D}, W \cup \overline{P},$
$\overline{B} \cup \overline{D}, \overline{B} \cup \overline{P}, \overline{B} \cup \overline{W}, \overline{D} \cup \overline{P}, \overline{D} \cup \overline{W}, \overline{P} \cup \overline{W}\, \rangle$

No property is selected during Phase 1, because Dog does not remove any distractors. During Phase 2, one property after another is rejected. The first property that is true of all elements of S while also removing distractors is $P \cup W$. This property happens to remove all distractors at once, causing the algorithm to end with $poodle \cup white$. If we modify the example by letting $[\![black]\!] = \{a, c\}$ (rather than $\{a, b, c\}$) and $S = \{b, c, d, e\}$ (rather than $\{a, b, d, e\}$), then the Logical Form $\overline{black} \cup \overline{poodle}$ is found.

The algorithm $\text{IA}_{\text{Boolean}}$ is not only incremental *within* a phase, but also from one phase to the next. Once a property has been selected, it will not be abandoned even if properties selected during later phases make it logically superfluous. As a result, one may generate Logical Forms like $X \cap (Y \cup Z)$ (e.g., "white (cats and dogs)") where $Y \cup Z$ (e.g., "cats and dogs") would have sufficed (because $(Y \cup Z) \subseteq X$). This is not unlike the redundancies generated by the original Incremental Algorithm, but more dramatic. Adaptations could be made. For instance, phases might run separately before running in combination: first (as usual) Phase 1, then 2, then (as usual) 1&2, then 3, then 1&3, then 2&3, then (as usual) 1&2&3, etc. As a result of this adaptation, the Logical Form $Y \cup Z$ would be generated on account of Phase 2 alone.

Double incrementality does not save $\text{IA}_{\text{Boolean}}$ from intractability. To estimate running time as a function of the number of properties (n_a) in the Knowledge Base and those in the Logical Form (n_l), one can mirror an argument in Dale and Reiter (1995, section 3.1.1) to show that the maximal number of properties considered equals

$$\sum_{i=1}^{n_l} 2 \binom{n_a}{i} = \sum_{i=1}^{n_l} 2 \frac{n_a!}{i!(n_a - i)!}$$

(The factor of 2 derives from inspecting each atom and its negation.) If $n_l << n_a$, then this is in the order of $n_a^{n_l}$. The algorithm can be pruned to make it polynomial. By cutting off after Phase 1, for example, only literals would be combined.completeness would be lost, but only for references to non-singleton sets, because set union does not add descriptive power where the description of singletons is concerned. The number of properties to be considered by this simpler algorithm equals $n_a^2 + 2n_a - 1$. To produce Logical Forms like $White \cap (Cat \cup Dog)$, the algorithm could be cut off one phase later, leading to a worst-case running time of $O(n_a^3)$, and so on.

Earlier, we proved Intersective Completeness for two versions of Dale and Reiter's Incremental Algorithm. We now prove Boolean Completeness for $\text{IA}_{\text{Boolean}}$, the Boolean extension of $\text{IA}_{\text{Plural}}$.

Theorem 4 Completeness of $\text{IA}_{\text{Boolean}}$. Assume there are at most denumerably many properties, and finitely many distractors (one or more). Then if a set can be individuated distributively by any Boolean combination of properties, then $\text{IA}_{\text{Boolean}}$ will find such a combination. *Proof*:
Any Boolean expression can be written in Conjunctive Normal Form (CNF), that is, as an intersection of unions of literals (e.g., Fitting 1996). The theorem follows from the following Lemma.

Lemma 1 Let φ be a CNF whose longest union has a length n (conjoining n literals). Then $\text{IA}_{\text{Boolean}}$ will find a Logical Form φ' that is coextensive with φ, in at most n phases. This is proven by induction on the size of n.
Base step: If $n = 1$, the Lemma is equivalent to completeness of $\text{IA}_{\text{Plural}}$, whose proof is analogous to that of the completeness of IA, replacing $\{r\}$ by S.
Induction step: Suppose the Lemma is true for all $n < i$. Now consider a CNF φ whose longest union has length i; let φ contain m unions of length i, namely, $\varphi_1 \cap .. \cap \varphi_m$. Then φ can be written as the CNF $\chi \cap \varphi_1 \cap .. \cap \varphi_m$, where all the unions in χ have length $< i$. The Lemma is true for all $n < i$, so if χ is sent to $\text{IA}_{\text{Boolean}}$, then the output is some χ' such that $[\![\chi']\!] = [\![\chi]\!]$, in fewer than i phases; so if, instead, φ is sent to $\text{IA}_{\text{Boolean}}$, then, after $i - 1$ phases, some possibly incomplete Logical Form η has been found, such that $[\![\eta]\!] \subseteq [\![\chi]\!]$. Also, $[\![\varphi]\!] \subseteq [\![\eta]\!]$. Phase i inspects all unions of length i, including each of $\varphi_1, .., \varphi_m$. Therefore, unless a Logical Form coextensive with φ is found before Phase i, one will be found during Phase i. To see this, suppose the algorithm finds ψ such that $[\![\psi]\!] = [\![\varphi_1]\!] \cap .. \cap [\![\varphi_m]\!]$, then $[\![\chi]\!] \cap [\![\psi]\!] = [\![\varphi]\!]$; but $[\![\varphi]\!] \subseteq [\![\eta]\!] \subseteq [\![\chi]\!]$, therefore also $[\![\eta]\!] \cap [\![\psi]\!] = [\![\varphi]\!]$. □

8.5 Optimization of Generated REs

REG algorithms sometimes generate Logical Forms in two steps: the first step uses a simple method to produce a potentially lengthy Logical Form, whereas the second removes redundancies. For instance, Ehud Reiter proposed to generated a potentially overspecified initial Logical Form of a (single) target referent, after which properties that are not necessary for singling out the referent are removed from this initial Logical Form [Reiter, 1990b]. Two-step procedures are not unknown in psycholinguistics either – Levelt's famous model of language production (discussed in our section 1.4) allows the speaker to monitor her own speech and modify it on the basis of this self-monitoring ([Levelt, 1989], chapter 12). We shall encounter a different type of self-monitoring in chapter 12, where tentative descriptions are enhanced if they are likely to cause trouble to hearers.

A two-step procedure for generating references to sets was proposed in [van Deemter, 2002]. The idea was not merely to remove superfluous properties, but to re-structure the initial Logical Form completely, making sure that the resulting Logical Form is *logically equivalent* to the initial Logical Form, aiming for the shortest equivalent. For instance, a Logical Form of the form $(p \cup q) \cap (p \cup r)$ is simplified to $p \cup (q \cap r)$. This procedure reduce the size of many REs, and can be implemented using algorithms that are used routinely in the design of logic circuits (e.g., [McCluskey, 1965]). The proposal of [van Deemter and Halldórsson, 2001] is open to the same improvements.

These optimizations are useful but they do not always lead to the shortest RE possible, because they only look at logical equivalence. In other words, they only replace an initial Logical Form with a shorter one if the two Logical Forms corefer in *all* possible domains, whereas one may want to do the same whenever they happen to corefer *in the domain* to which they are applied.

For example, suppose the initial Logical Form is as above, but with p' replacing the second occurrence of p, that is, $(p \cup q) \cap (p' \cup r)$. Now suppose p and p' are co-extensive, because they hold true of the same entities in the domain at hand. For example, the domain consists of the people in a room. Now p might say *left-handed* and p' might say *wearing a kilt*, and the two people in the room who are left-handed are also the only ones wearing a kilt. The same optimization is possible as before, but the program will fail to find it, because in other domains, p and p' may not be co-extensive. The problem can also apply to combinations of properties, for example when $p \cup q$ has the same extension as

$p' \cup q'$. Gardent's constraint-based approach would solve the problem, always producing the shortest RE possible (section 6.4). However, given what is known about human reference production (e.g., chapter 5), it seems unlikely that the shortest REs are always the ones most likely to be produced, or the ones that are most effective for a hearer. What is needed is a systematic investigation into speakers' strategies for referring to sets, analogous to earlier investigations into the classic REG task.

8.6 Issues Raised by the Algorithms Proposed

The algorithm of section 8.4 has given rise to a number of questions. Claire Gardent, for example, drew attention to situations where earlier proposals produced very lengthy Logical Forms, and proposed a reformulation of REG as a Constraint Satisfaction problem, which finds optimally brief references to sets (see section 6.4). Helmut Horacek has argued that it is generally better to generate Logical Forms in Disjunctive Normal Form (DNF; unions of intersections of literals) than the CNF-based Logical Forms generated by [van Deemter, 2002]. Horacek therefore proposed an algorithm that first generates Logical Forms in CNF and then convert these into DNF, skipping superfluous disjuncts [Horacek, 2004]. To see how this can work, let's modify our old example again, assuming that all dogs are either poodles or alsatians:

TYPE: Dog ($\{a, b, c, d, e\}$), Poodle ($\{a, b\}$), Alsatian ($\{c, d, e\}$)
COLOUR: Black ($\{a, c\}$), White ($\{b, e\}$)

To refer to $\{b, c\}$, Horacek would start with an CNF Logical Form such as $(white \cup alsatian) \cap (black \cup poodle)$ ("the dogs that are white or alsatians, and also black or poodles"). This would be converted into DNF: $(white \cap poodle) \cup (white \cap black) \cup (alsatian \cap poodle) \cup (alsatian \cap black)$ after which the middle two disjuncts would be dropped, because nothing is both white and black, or both an alsatian and a poodle. The outcome can be worded as "the white poodle and the black alsatian", which looks like a substantial improvement, regardless of whether the aim of REG is humanlikeness or benefit for hearers. Later work has tended to agree with Horacek in opting for DNF instead of CNF [van Deemter and Krahmer, 2007], [Gatt, 2007], [Khan et al., 2012]. In the remainder of this chapter, we shall do the same, focussing on conjoined NPs ("the so-and-so and the so-and-so") in particular.

We have discussed algorithms that extend the referential power of REG, enabling the generation of distinguishing descriptions where this was not previously possible. This gives rise to a range of new questions. For example, it is difficult to see how the REs discussed in the previous sections can be generated in reasonable time, raising the question of how human speakers are able to do this; algorithmic complexity has always been an issue in relation of the classic REG problem (see section 4.8), but in the new situation it becomes huge.

It is difficult to choose between different distinguishing descriptions that contain combinations of Boolean operators. In the words of Fitzgerald and colleagues, *"For a target set of objects, the number of logical forms that can be used to describe it grows combinatorially with the number of observable properties, such as color and shape. However, only a tiny fraction of these possibilities are ever actually used by people"* [FitzGerald et al., 2013]. Some important issues were raised by Gardent and Horacek, as we have seen, and more recently Machine Learning approaches have been applied with considerable success, as by Fitzgerald and colleagues, who sought to learn the corpus-based likelihood of each logical form using a new method (based on stochastic gradient descent). Additionally, there is the question of how Boolean operators should be combined with other phenomena: suppose a REG algorithm was able to use relational descriptions as well as negation. Now what is preferable: adding a relational property ("... in the wooden shed"), or adding a negated property ("... which is not a poodle")? Empirical investigations analogous to those of chapter 5 are called for.

We leave these issues here, devoting the remainder of this chapter to two other thorny issues: the internal coherence of disjunctive references and the risk that surface ambiguities may break the clarity that REG algorithms are designed to achieve. The strategic question of how much effort the research community should invest in the investigation of the types of complex REs that this chapter is beginning to explore is discussed in section 10.7.

8.7 Lexical Coherence in Conjoined REs

So far, when we are seeking to generate a conjoined RE, we are essentially splitting the referent set into two or more parts, each of which is referred to separately, as when we say "the white poodle and the black alsatian", for instance. We are pretending to deal with two independent problems even though, in reality, the two problems are connected.

IDENTIFIER	SPECIES	ORIGIN
a	*lion*	*Kenya*
b	*lion*	*China*
c	*tiger*	*China*
d	*elephant*	*India*

Table 8.1
How to refer coherently to the set $\{a, c\}$*?*

To see why this is a problem, let us visit the infirmary of our zoo again, which was introduced in chapter 1. An elephant is added to the casualties, and certain facts (such as the weight of an animal) are left out of consideration (Table 8.1). Let us look at some of the subsets of this domain, asking how to refer to them. The set $\{c, d\}$ can be referred to briefly and uniformly as "the tiger and the elephant", because there is only one of each. The set $\{b, c, d\}$ might be referred to as "the Asian animals" (provided the algorithm knows where each country is located). What, however, is the best way to refer to $\{a, c\}$? A number of options suggest themselves, including

1. The Kenyan animal and the tiger.
2. The Kenyan lion and the Chinese tiger.

The first of these lacks *coherence*, because it describes the two animals using different conceptual perspectives. The second is more uniform, but more verbose as well. It appears that either coherence or brevity has to be sacrificed. How to choose?

A key question is whether these issues are best viewed at a conceptual level or a lexical one. A *conceptual* view is offered in [Aloni, 2002] (see our chapter 2, section 2.2), but this solution was rather rigid, implying that reference always needs to stick with one conceptual perspective (e.g., an animal's country of origin), which is not always possible. A *lexical* view could exploit existing methods in NLG, based on n-gram-based filtering (e.g., [Langkilde and Knight, 1998]): the idea would be to generate a large number of conjoined NPs and to select the one that resembles the expressions in a corpus most, in terms of its bi-grams or tri-grams. Filtering by n-grams, however, is ill suited for addressing choices of words that are more than 2-3 words apart, and data sparsity can easily become a problem. It appears that a different approach is required.

Albert Gatt and I wanted to know whether a systematic perspective can be brought to these issues [Gatt and van Deemter, 2007]. Gatt had the starting intuition that, for an RE of the form "the so-and-so and the so-and-so" to be coherent, the words – and especially the nouns – in the RE need to be as similar to each other as possible, taking into account that complete coherence (i.e., where all the words in a description are taken from the same conceptual perspective) may not be compatible with the aim of referring uniquely. The key concepts in this story are coherence and similarity.

Similarity had been studied extensively by Dekang Lin [Lin, 1998b], [Lin, 1998a], building on existing Information Theoretical methods for computing the information contained in a statement (e.g., [Fano, 1961], [Cover and Thomas, 1991], see also chapter 13). Lin had argued that the similarity of two arbitrary entities A and B should be measured by comparing the amount of information needed to describe what A and B have in common with the information needed to fully describe what A and B are.

Focussing on distributional similarity, Lin based the description of a word on the set of dependency triples $\langle rel, w, w' \rangle$ found in a corpus, where rel is a grammatical relation, w the word of interest, and w' its co-argument in rel (or conversely if w' is the word of interest). For instance, some of the triples associated with the word "master", obtained from the British National Corpus, are ⟨subject-of, "master", "attend"⟩, and ⟨subject-of, "master", "write"⟩.

Lin viewed a grammatical triple $\langle rel, w, w' \rangle$ as a *feature* of both w and w' (e.g., "master" has the feature Subject-of("attend")), allowing a word w to be described by a set of features $F(w)$ that characterize its syntactic behaviour. Using this idea, the information that A and B have in common is $F(A) \cap F(B)$. Lin went on to formalize the amount of similarity between two words in the following formula (reminiscent of the Dice metric of chapter 5), where I stands for the amount of information needed to describe a set of features:

$$sim(w_1, w_2) = \frac{2 * I(F(w_1) \cap F(w_2))}{I(F(w_1) + I(F(w_2)))}, \tag{8.1}$$

Following [Lin, 1998a], the definition of $I(F(w))$ takes into account a number of factors, including the overall frequency of each of the relations with which w occurs in the corpus. Gatt obtained F and I values from SketchEngine2 [Kilgarriff, 2003], which contains information about word similarity and the mutual information of grammatical triples, based on estimates from the British National Corpus. Similarity between pairs of nouns was estimated on the basis

distributional sim	ontological sim	EXAMPLE
high	*high*	*the leader and the chairman*
high	*low*	*the manager and the council*
low	*high*	*the department and the resource*
low	*low*	*the garden and the police*

Table 8.2

Two types of similarity (sim) *pitted against each other: examples of* REs *used as materials for one of the experiments in [Gatt and van Deemter, 2007].*

of the three grammatical relations of (a) the likelihood of two nouns occurring as subjects of the same verb, (b) the likelihood of two nouns occurring as objects of the same verb, and (c) the likelihood of two nouns being pre- or post-modified by the same adjectives.

Gatt carried out a number of experiments to assess whether this notion of distributional similarity (which formalizes the degree to which two words display the same *syntactic* behaviour) is a better predictor of the quality of a conjoined noun phrase than other kinds of similarity. These experiments suggest an affirmative answer. In an experiment based on magnitude estimation, for instance, participants were shown REs of the form "the noun_1 and the noun_2" that varied in terms of both their ontological similarity and their distributional similarity (Table 8.2). The former was measured in terms of the distance between noun_1 and noun_2 in terms of the number of edges between concepts in the WordNet IS-A nominal hierarchy [Pedersen et al., 2004]; the latter was measured using Lin's approach. Participants were asked to indicate, by moving a slider, for each of a large number of REs, how acceptable they found this RE. It turned out that participants' acceptability judgments lined up much better with distributional similarity than with ontological similarity.

Based on this and other experiments, Gatt designed a complex algorithm that, given a domain and a referent set containing two or more elements, generates expressions that refer to this set and whose words are as distributionally similar to each other as possible. Rather than testing the algorithm in the manner of chapter 5, we decided to test a specific prediction inherent in it, namely, that distributional similarity (i.e., *coherence*) can trump brevity.

The evaluation compared readers' preference for descriptions that were optimally brief or not ($+/-b$) and also either optimally coherent or not ($+/-c$). Non-brief descriptions took the form *the A, the B and the C*. Brief descriptions aggregated two disjuncts into one (e.g., *the A and the Ds*, where D comprises

the union of B and C). Materials offered to participants consisted of specially constructed items such as the following:

Three old manuscripts were auctioned at Sothebys:

(1) *One of them is a book, a biography of a composer.*
(2) *The second, a sailor's journal, was published in the form of a pamphlet. It is a record of a voyage.*
(3) *The third, another pamphlet, is an essay by Hume.*

Continuations:
$(+c, -b)$ *The biography, the journal and the essay were sold to a collector.*
$(+c, +b)$ *The book and the pamphlets were sold to a collector.*
$(-c, +b)$ *The biography and the pamphlets were sold to a collector.*
$(-c, -b)$ *The book, the record and the essay were sold to a collector.*

In a forced-choice task, participants were asked which continuation they found *most natural*. We had expected to find that $+c$ descriptions are preferred over $-c$, that $(+c, -b)$ descriptions are preferred over ones that are $(-c, +b)$ (i.e., coherence is more important than brevity), and that $+b$ descriptions are preferred over $-b$. We found that the first two hypotheses were confirmed, but the last one was not. Not only were coherent descriptions preferred over non-coherent ones when the brevity level $(+/-b)$ was kept constant, but where there was a trade-off between brevity and coherence, participants were much more likely to select a coherent description than a brief one). To our surprise, brevity on its own (i.e., the difference between conjunctions of 2 or 3 NPs) did not have any measurable effect.

It is possible that an effect of brevity would have been found if the differences between $+b$ and $-b$ had been magnified. Be this as it may, the fact that coherence, formalized in the corpus-based manner of Lin, was more important than brevity suggests that it is a factor of genuine relevance for the quality of a conjoined RE.

It does not seem farfetched to conjecture that lexical coherence may work across larger stretches of discourse as well, as when a number of referents are discussed sequentially, for example, even when the REs in question are separated by considerable amounts of text. In fact, it is difficult to see why the effects of lexical coherence should be limited to REs: the acceptability of indefinite NPs ("a book, a record and an essay") and quantified NPs ("all books, records and essays") is likely to be affected by the same factors, for example.

If this is true, then the lexical coherence effect has potential relevance to many areas of language.

The methods discussed above appear to have the potential to shed light on some of the hardest questions in REG, involving the Gricean notion of Relation. It would be possible, for instance, to use lexical similarity to favour the selection of properties (or words) that are most similar, on average, to the words in the previous few sentences (or the semantic content that these sentences are trying to express), thereby weeding out some of the *irrelevant* REs discussed by Kronfeld (section 4.2). Alternatively, one could use ontological similarity to achieve the same. It would be interesting to see how the different similarity-based approaches to relevance affect the readability and intelligibility of a text.

8.8 Avoiding Surface Ambiguities

The main point of the algorithms discussed so far is to individuate a target, making sure that nothing else answers to the properties accumulated by the algorithm than the intended referent. Yet all this work can come to naught if later modules realize a Logical Form as an ambiguous string of words.

Perhaps the most obvious situation in which surface ambiguities can compromise REG results from *lexical* ambiguity. To borrow an example from Siddharthan and Copestake's [Siddharthan and Copestake, 2004], suppose we want to refer to an individual in a room full of royalty; we may want to let our RE use the fact that our referent is the oldest of the kings in the room, and express this by saying "the old king". But if the *previous* king is also in the room, then this RE is referentially ambiguous. It is true that human hearers will often interpret an RE *charitably*, so if there are 2 previous kings in the room, neither of whom is the oldest king in the room, then hearers will tend to understand that "old" here is about age, because otherwise the NP would fail to refer uniquely. But if there is only one previous king in the room, then the RE can be understood as referring in two different ways, and confusion can result, which is something generation programs should normally try to prevent.

Matters become more complicated when ambiguity is caused by *syntactic* structure. To see what's at stake, let us turn to an example in [Khan et al., 2012], where scopally ambiguous conjoined NPs take centre stage. Consider a meadow with various animals, and suppose the generator's task is to single out the black sheep and black goats. Suppose a REG algorithm has generated the Logical Form $(black \cap sheep) \cup (black \cap goats)$.

This could be realized as: "the black sheep and the black goats" or "the black sheep and goats". The former is structurally unambiguous but lengthy and arguably disfluent; the latter is ambiguous between $(black \cap sheep)$ (with narrow scope for the adjective) and $black \cap (sheep \cup goats)$ (with wide scope for the adjective); only the latter is equivalent to the intended Logical Form.

So once again, brevity is coming up against another factor: in the previous section, we discussed tensions between brevity and coherence; in the present case, we are facing a tension between brevity and avoidance of ambiguity. The question is how to balance these two potentially conflicting factors.

A key insight is that not all "theoretical" ambiguities need to be avoided, because not all ambiguities that a parser can find are likely to lead to actual confusion. For although most sentences in a corpus are multiply ambiguous, most parses are hugely unlikely.

Imtiaz Khan, Graeme Ritchie, and I hypothesized that the corpus-based techniques that had helped Gatt to assess the coherence of a conjoined NP could also help us to predict the likelihood of a parse. Building on [Chantree et al., 2005], Khan saw expressions of the form "the Adj Noun1 and Noun2" as subject to two competing forces: a Coordination Force, whereby Noun1 and Noun2 attract each other to form a syntactic unit, and a Modification Force, whereby Adj and Noun1 attract each other.

We speak of Strong Coordination Force (SCF) if the collocational frequency between the two nouns is high, and of Weak Coordination Force (WCF) if the collocational frequency is low. Similarly, we speak of Strong Modification Force (SMF) if the collocational frequency of Adj is high with Noun1 and low with Noun2, and a Weak Modification Force (WMF) otherwise. Khan decided to define high collocational frequency between two words as a situation in which either of the two words appears among the top 30% collocates of the other word in the grammatical relation of interest; low collocational frequency was defined as a situation in which neither of the two words appears among the top 70% collocates of the other word in the grammatical relation. Between 30% and 70%, frequency is neither low nor high, so some phrases manifest neither strong nor weak Coordination/Modification Force. Kilgarriff's SketchEngine2, with the BNC as a corpus, was employed to obtain frequencies. Four hypotheses were formulated:

Hypothesis 1: If SCF and SMF, a narrow-scope reading is most likely.
Hypothesis 2: If SCF and WMF, a wide-scope reading is most likely.

	Force	Predicted	Judgments	p-Value
Hypothesis 1	SCF & SMF	NS	NS (51/60)	$< .001$
Hypothesis 2	SCF & WMF	WS	WS (55/60)	$< .001$
Hypothesis 3	WCF & SMF	NS	NS (46/60)	$< .001$
Hypothesis 4	WCF & WMF	WS	WS (54/60)	$< .001$

Table 8.3

Results of one of the experiments in [Khan et al., 2012], which saw all four hypotheses confirmed. WS *is wide scope,* NS *is narrow scope,* SMF *is Strong Modification Force, and* WMF *is Weak Modification Force. p-Values are based on a one-tailed sign binomial test. The experiment suggested that corpus-based frequencies can offer a good mechanism for predicting the preferred reading of a scopally ambiguous conjoined* NP.

Hypothesis 3: If WCF and SMF, a narrow-scope reading is most likely.
Hypothesis 4: If WCF and a WMF, a wide-scope reading is most likely.

Hypotheses 2 and 3 are intuitively obvious, because both forces operate in the same direction; the other two were based on preliminary studies. As it happened, all four hypotheses were confirmed, as summarized in Table 8.3. These findings can be summarized using the following Prediction Rules, which make Modifying Force the decisive factor:

1. If WMF, then WS
2. If SMF, then NS

Based on a number of experiments, Khan knew that if readable and intelligible sentences are to be generated, then the risk of misunderstandings needs to be balanced against verbosity. For example, if two sentences are equally clear and one is much shorter than the other, then, *ceteris paribus*, the shorter one is read and understood more quickly [Khan et al., 2012]. For this reason, Khan's REG algorithm balances brevity and clarity.

A crucial notion in the algorithm is that of a "clear" RE, whose definition is shown in Figure 8.4. The notion of a "brief" RE is also used in a specially defined sense: descriptions of the form "the Adj Noun1 and Noun2" count as brief; descriptions of the form "the Adj Noun1 and the Noun2" (for narrow scope) and "the Adj Noun1 and the Adj Noun2" (for wide scope) are non-brief. That is, brevity has a specialized sense involving the presence/absence of the determiner (the), and possibly an Adj before the second noun. The non-brief expressions are always syntactically unambiguous, but the brief NPs are syntactically ambiguous. Crucially, syntactically ambiguous NPs can be clear.

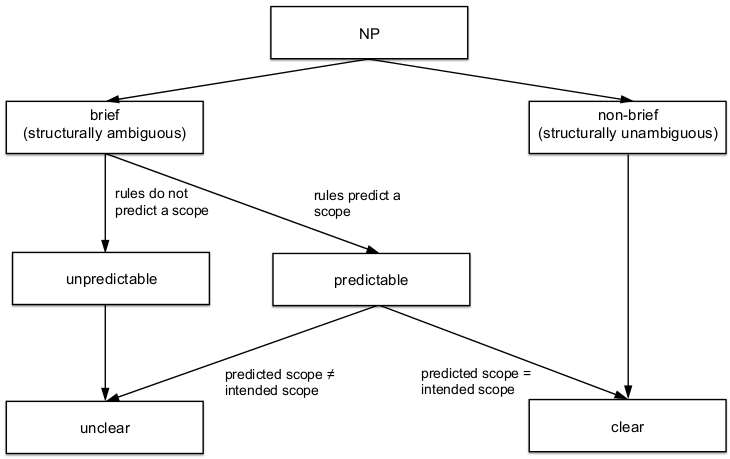


Figure 8.4

This diagram summarizes how the REG *algorithm of [Khan et al., 2012] decided whether to classify a conjoined* NP *as "clear" or "unclear" (as these notions are defined in the text). The algorithm avoids unclear expressions whenever it can. Note that structurally ambiguous* NP*s can be clear. Not only those expressions are unclear for which the corpus-based method fail to predict a reading, but also those for which this method predicts an unintended reading.*

Note that not only those expressions are unclear for which the corpus-based method *fail* to predict a reading, but also those for which this method predicts a reading that does not correspond with the input to the generator.

The generation algorithm starts by generating formulas in Disjunctive Normal Form (DNF). The algorithm, which builds on [Gatt, 2007], partitions the referent set into non-overlapping subsets and builds an initial Logical Form denoting each subset using a conjunction of properties. The formulas that matter, from the present point of view, are of the form $(A \cap N_1) \cup (A \cap N_2)$. The formulas are like conventional set theoretic expressions, but their building blocks are English words rather than names of sets. For example, we use formulas such as $man \cup (big \cap dog)$ to denote the set of domain objects that are either men or big dogs. The algorithm consists of five stages and is exemplified below. Once an initial Logical Form is produced, this Logical Form is

transformed to produce a variety of other formulas, each of which is logically equivalent to the initial formula. The following transformation rules were used:

a. $(A \cap N_1) \cup (A \cap N_2) \Rightarrow A \cap (N_1 \cup N_2)$
b. $X \cup Y \Rightarrow Y \cup X$

These rules apply as often as possible, producing new formulas. Then during Linguistic Realization, these formulas are converted into strings of words. An example of a Realization rule is $(Adj \cap Noun1) \cup (Adj \cap Noun2) \rightarrow$ *the Adj Noun1 and the Adj Noun2*, which converts a formula directly into English. Realization produces a set of NPs, each of which could be used to refer to the target set. The problem now is to select the best NP, which is done during the last two stages, where the results of the experiments are utilized. First, the algorithm assesses the clarity and brevity of all the NPs that are available at this point. This stage makes use of the Prediction Rules, which stated that WMF $\rightarrow$ WS and SMF $\rightarrow$ NS. During its selection phase, the algorithm prefers clear NPs over unclear ones; and if several NPs are clear, then the choice between them is made on the basis of brevity.

Algorithm 15 Khan's algorithm for managing surface ambiguities

Input: A domain of objects, containing a non-empty target set S consisting of elements of the domain; also including a non-empty set M of distractors (i.e., domain elements that are not elements of S). A set $\mathcal{P}$ of properties true of each element of S. Brevity (line 4) is defined in the text. Clarity (line 4) is defined in Figure 8.4.
Output: An NP whose extension denotes the target referent set S, if such an NP exists. The NP embodies an experimentally motivated compromise between clarity and brevity.

1: Construct a Logical Form $\mathcal{D}$ in Disjunctive Normal Form
2: Use transformation rules to obtain a set of Logical Forms equivalent to $\mathcal{D}$
3: Perform Linguistic Realization on each of these Logical Forms;
   the result is a set N of NPs
4: Assess the clarity and brevity of each element of N
5: **if** some forms in N are clear and others are not **then**
6: remove from N all those forms that are not clear
7: **if** some remaining forms in N are shorter than others in N **then**
8: remove from N all those forms that are shorter than others in N
9: From the remaining forms in N make an arbitrary choice

The following example illustrates Khan's algorithm, writing R for "radical", S for "student", and T for "teacher". Suppose the Initial Logical Form (produced by line 1) is (a) $(R \cap S) \cup (R \cap T)$ (radical students and radical teachers). Transformation (line 2) produces three additional Logical Forms:

(b) $(R \cap T) \cup (R \cap S)$, (c) $R \cap (S \cup T)$, (d) $R \cap (T \cup S)$. Linguistic Realization (3) of these Logical Forms results in the following four NPs: (a) "The radical students and the radical teachers", (b) "The radical teachers and the radical students", (c) "The radical students and teachers", and (d) "The radical teachers and students". Each of these NPs is then tested for clarity, making use of the Prediction Rules.

The Prediction Rules predict wide scope for (d), because the relation between radical and teachers is WMF (according to the BNC data). The Initial Logical Form was a wide-scope interpretation for this phrase, so despite its theoretical ambiguity, (d) is judged to be clear. The rules do not predict a scope for (c), because the relation between radical and students is neither WMF nor SMF; hence, (c) is judged to be unclear. The other two NPs, (a) and (b), are clear (because unambiguous) but less brief than (d). The NPs (a) and (b) are less brief than (c) and (d). Based on these assessments, the NP (d) is generated, which is clear and brief. The algorithm has opted for a brief noun phrase. In other cases, where the Prediction Rules pan out differently, a more elaborate noun phrase is generated. An example is "radical students and radical soldiers", where omission of the second adjective would create misunderstandings. The contrast between these examples shows that, where syntactic ambiguity is involved, words can make all the difference.

This study teaches us an important lesson: identification of the intended referent cannot be guaranteed by an approach to generation that only pays attention to logic: the generation algorithm has to be aware of syntactic ambiguities as well, and of the probability of misinterpretation. Luckily, there is now a wealth of corpus-based and experimental work to inform this probability. Note that the evidence, in this case, does not tell us how speakers behave; it only tells us how they should behave if they wanted to produce sentences that are read and understood quickly. The status of a computational model, like Khan's, that aims to maximize utility, comes up in sections 1.5 and 16.1 (where the first dimension of variation between models is discussed).

8.9 Beyond Sets of Objects

There is more to the topic of this chapter: all we have done is present some basic computational models for referring to sets, and discuss ways in which these models can handle lexical coherence and syntactic ambiguity. Reference to sets poses many other challenges, some of which will be explored

in chapter 10, where we shall be studying complex REs such as "the men who feed two dogs". An interesting issue that we will not discuss is the choice between different determiners, which is much more difficult in connection with plurals (where options may include "each", "every", "both", and so on [Shaw and McKeown, 2000]) than in connection with singulars.

Logically, the move from individuals to sets can be carried further. We often refer to entities that are neither individual objects nor sets of individual objects. We can say things like *water* and *yellow sand* in order to refer to what is sometimes known as "stuff". With few exceptions (e.g., [Dale, 1989b], [Dale, 1988], [Mitchell et al., 2012]), the Computational Linguistics community has not yet touched this area. In chapter 14, we shall encounter an approach to reference to *geographical areas* that reduces a geographical area to a large but finite *set* of points on a map, bringing them into the orbit of the present chapter [Turner et al., 2008]. First, however, we need to discuss other ways in which classic REG models have been extended.

8.10 Summary of the Chapter

Sets are targets for reference as are individual objects, yet their generation raises many issues that do not affect the classic REG problem. This chapter has outlined some basic algorithmic approaches that can be taken, and discussed a few of the difficult research questions thrown up by reference to sets.

- The logic of sets is complex, and this complexity affects the art of referring to sets. For example, reference to sets using collective properties is a little explored area that is much in need of further study. [Section 8.1]
- In the simplest cases, where the elements are bound together by a property that they all share (and that is not shared by any distractors), sets can be referred to using the classic REG algorithms, as long as a suitable TYPE (i.e., a suitable noun) can be found that covers all elements of the target set. [Section 8.1]
- Negation can be a crucial operation, not only when referring to sets, but even when referring to a single referent. [Section 8.2]
- Disjunctive REs have no role to play in referring to a single object, but when the aim is to refer to a set, disjunctive references (as when we use the Logical Form $Horse \cup Cow$, "the horses and cows") are hard to do without. [Section 8.2]

- Satellite sets can be employed as a component of REG programs. They can also help us to compute quickly which sets, or which objects, can be identified uniquely given a certain set of properties. [Section 8.3]
- Disjunctive REs become more acceptable if they display a certain thematic or lexical coherence. For example, it is better *ceteris paribus* to refer to a set as "the book and the pamphlets" than "the biography and the pamphlets". The precise nature of this coherence is gradually becoming clear. Crucially, the most coherent RE may not be the best one in other respects; for example, it may be lengthier than others. [Section 8.7]
- Disjunctive REs can give rise to syntactic ambiguities. Investigation into these ambiguities suggests that it is not always a good idea to avoid them, given that the avoidance of ambiguity can create other infelicities: it seems preferable to avoid only those ambiguities that are likely to lead to actual misunderstanding. Corpus-based probabilities can help to predict which ambiguities fall into this category. [Section 8.8]
- Transformations on Logical Forms, where the transformation respects logical equivalence, can simplify an initial Logical Form or broaden the set of REs to be considered by a REG algorithm. [Section 8.8]

9 Third Extension: Using Gradable Properties

We have so far avoided some of the main difficulties that language poses to theorists and modellers. For example, we have avoided context dependence and vagueness. In this chapter we shall see that models of reference production become more complex when these issues are taken into account. Specifically, we shall study the role of gradable properties in reference production.

Gradable properties have borderline cases, in which it is unclear whether the words in question can be applied. Consider the word "giant": a height of 240 cm makes you a giant; a height of 170 cm doesn't, but somewhere in between, the matter can be unclear. We noted something similar in our opening chapter when we wondered whether the tiger in Figure 1.1 could be called "huge". Properties or expressions that have borderline cases are called vague.

It might be thought that referring expressions (REs) cannot contain gradable properties, because their vagueness jeopardizes the aim of identifying a referent. Yet REs often contain gradable words. Following Manfred Pinkal, REs that contain gradable expressions will be called *vague descriptions* [Pinkal, 1979], even in situations where the expression as a whole is clear (i.e., not vague). Gradable expressions are context dependent: to be a small elephant is not to be small and to be an elephant, but to be "small for an elephant". Context dependence will complicate REG considerably.

To see how vague descriptions can work, suppose you enter a vet's surgery with two dogs of different sizes. The vet asks "Who's the patient?", and you answer "the big dog". This answer will allow the vet to pick out the patient just as reliably as if you had said "the one on the leash"; the vagueness of "big" is irrelevant. This shows how potentially *vague* properties can contribute to the *precise* task of identifying a referent. Additional detail (e.g., about the size of the dog) does not improve identification, and might even detract from it, because measurement is more error prone than comparison [Lipman, 2009], [van Deemter, 2009a], [van Deemter, 2010].

Let's see to what extent computational models of reference can capture the peculiarities of vague descriptions.[1] We start by discussing the semantics and pragmatics of vague descriptions (sections 9.1 and 9.2), after which we briefly examine the experimental evidence (section 9.3). In section 9.4 we discuss one way in which REG algorithms can determine the content of vague descriptions, in the plural (cf., chapter 8) as well as the singular, as long as only one gradable dimension is involved: once there are several, things become problematic.

1 This chapter integrates and extends material from [van Deemter, 2000], [van Deemter, 2006], [Mitchell et al., 2010], [Mitchell et al., 2011b], and [Mitchell et al., 2013a].

After a discussion of its implications for the idea of Incremental REG, we turn to a concrete case study to see what challenges face us when gradable adjectives are used for talking about real-world objects (section 9.6). We conclude the chapter with a discussion of *salience* – a core concept in communication, as we have seen – as a gradable dimension (section 9.7), and its implications for the risk that referring expressions may be misunderstood.

9.1 The Semantics of Vague Descriptions

Before turning to reference, ket's examine the various forms vague adjectives can take, using the word "large" as an example: they may or may not contain a numeral n (positioned before or after the adjective), and the gradable adjective (Adj) may be in base ("large") or superlative form ("largest"):

1. *The* (n) *Adj(est) N* (e.g., "the 3 largest mice")
2. *The Adj(est)* (n) *N* (e.g., "the largest 3 mice").

If *Adj* is in base form, we shall focus on the word order (1); if *Adj* is a superlative, then we focus on (2). We are limiting ourselves to *referential* uses, excluding cases like "This may be *the largest mouse in the world*", in which the expression ascribes a property to an already-identified object.

Different analyses are possible of what it means to be large: larger than average, larger than most, larger than some given baseline, and so on. It is doubtful that any one analysis makes sense for all definite descriptions. Consider a domain of three mice, sized 5 cm, 8 cm, and 10 cm. Here one can say

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

Clearly then, what it takes for the adjective to be applicable has not been cast in stone but is, at least to an extent, open to *fiat*: the speaker may decide that 8 cm is enough, or she may set the standards higher (cf., [Kennedy, 1999]). The numeral (whether it is implicit, as in (3), or explicit, as in (4)) enables the reader to draw inferences about the standards employed [Kyburg and Morreau, 2000], [DeVault and Stone, 2004]: (3), for example, implies a standard that counts 10 cm as large and 8 cm as not large. We shall ask how NPs like the ones in (3) and (4) can be generated, without asking how they constrain, and are constrained by, other uses of "large" and related words. We shall make the following simplification: in a definite description that expresses only properties that

are needed for singling out a referent, we take the base form of the adjective to be *semantically* equivalent with the superlative (and comparative) forms:

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

Although this is only meant as a rough approximation, a modest amount of empirical evidence for this position will be offered in section 9.3.

Clearly, the expression "Adj(est)" in (3) and (4) is context dependent. We simplify by pretending that the only contextually relevant factor is the comparison set: those elements of the noun denotation that are perceptually available. We disregard *functional* context-dependence, as when "the small hat" is the one too small to fit on your head. For the moment, we also disregard the global (i.e., not immediately available) context. For some adjectives, including the ones that Manfred Bierwisch called *evaluative*, this is clearly inadequate. Bierwisch argued that evaluative adjectives (such as "beautiful" and its antonym "ugly"; "smart" and its antonym "stupid", *etc*.) can be distinguished from the more frequent purely "dimensional" adjectives by the way in which they compare with their antonyms. For example, after [Bierwisch, 1989],

Hans is taller than Fritz $\Rightarrow$ Fritz is shorter than Hans.
Hans is smarter than Fritz $\not\Rightarrow$ Fritz is more stupid than Hans.

Bierwisch argued that the referent of an evaluative description should fall into the correct segment of the relevant dimension. (For Fritz to be "the stupid man", it is not enough for him to be the least intelligent male in the local context; he also has to be a fairly stupid specimen in his own right.)

Let us say more precisely what we will assume the different types of expressions to mean. For ease of reading, concrete examples (e.g., "large") will replace syntactic categories, but the analysis is meant to be general.

The largest *n* mouse/mice. The *n* large mice. Imagine a set C of contextually relevant animals. Then these NPs presuppose that there is a subset S of C that contains n elements, all of which are mice and such that (1) $C - S \neq \phi$ (i.e., not all elements of C are elements of S) and (2) every mouse in $C - S$ is smaller than every mouse in S. If such a set S exists, then the NP denotes S. The case where $n = 1$, realized as "The large(st) mouse", falls out automatically.

The large(st) mice. The above account can be extended to cover cases of the form *The [Adj]-(est) [N_{pl}]* (pl = plural), where the numeral n is suppressed: they will be taken to be ambiguous between all expressions *The [Adj]-(est) n [N]*, where $n > 1$. Sometimes this leaves only one possibility. For instance, in a domain where there are five mice, of sizes 4, 4, 4, 5, and 6 cm, the only possible value of n is 2, causing the NP to denote the two mice of 5 and 6 cm size.

Refinements of these ideas are discussed below.

9.2 Pragmatic Constraints on What Can Be Said

Models of language production have to do more than find a distinguishing description, (i.e., one that unambiguously denotes its referent): the description should also be felicitous (cf., section 1.5). In other chapters, we look empirically at different dimensions of felicity, such as humanlikeness (chapter 4) and benefits for human hearers (e.g., chapters 8 and 12). The present chapter will use broader brush stokes, aiming for REs that perform well in both respects.

Consider the question, discussed in the philosophy of language, whether it is legitimate, for a gradable adjective, to distinguish between observationally indifferent entities: Suppose two objects x and y are so similar that it is impossible to distinguish their sizes; can it ever be reasonable to say that x is large and y is not? A positive answer would not be psychologically plausible, because x and y are indistinguishable; but a negative answer would prohibit *any* binary distinction between objects that are large and objects that are not, given that it is always possible to find (or construct) objects x and y, one of which falls just below the divide, whereas the other falls just above it.

A production model can offer a subtle response: that the offending statement may be correct yet infelicitous.

Small Gaps. Expressions of the form *The (n) large [N]* are infelicitous when the gap between (1) the *smallest* element of the designated set S (henceforth, s^-) and (2) the *largest* N smaller than all elements of S (henceforth, s^+) is small in comparison with the gaps between the other elements [Thórisson, 1994], [Funakoshi et al., 2004], [Gorniak and Roy, 2004], [Fernández, 2009]. If this gap is so small as to make the difference between the sizes of s^- and s^+ impossible to perceive, then the expression is also infelicitous.

Dichotomy. When separating one single referent from one distractor, the comparative form is often said to be favoured ("Use the comparative form to compare *two* things"). We generalize this idea to situations where all the referents are of one size, and all the distractors of another.

Minimality. Unless *Small Gaps* and *Dichotomy* forbid it, we expected that preference should be given to the base form. In English, where the base form is morphologically simpler than the other two, this rule could be argued to follow from Gricean principles.

To keep matters simple, our algorithm will choose the base form if and only if the gap between s^- and s^+ surpasses a certain value, which is specified interactively by the user.

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

9.3 Empirical Grounding

Readers interested in experimental testing of the claims presented in sections 9.1 and 9.2 are referred to [van Deemter, 2004], [van Deemter, 2006], and [Barr et al., 2013]; the present section summarizes the main findings.

Despite the mitigating role of the Small Gaps constraint, and of Bierwisch's principle, a potential worry is that our algorithm might put too much emphasis on comparing the referent with the distractors, neglecting the question of whether the adjective is applicable to the referent. It might be thought odd, for instance, to describe a cup as "the tall cup" merely because it is taller than the other cups in the relevant domain. Suppose the referent cup is smaller than a *normal* cup and too small to drink from; can it still be called "the tall cup"?

As it happens, the literature suggests an affirmative answer to this question [Sedivy et al., 1999]. The authors asked subjects to identify the target of a vague description in a visual scene, for instance "the tall cup". The scene would contain three distractors: (1) a less tall object of the same type as the target (e.g., a cup that is less tall), (2) a different kind of object which previous studies had shown to be intermediate in height (e.g., a pitcher that is taller than both cups, but neither particularly short nor tall for a pitcher), and (3)

a different type of object to which the adjective is inapplicable (e.g., a door key). It did not matter much whether the adjective applied "intrinsically" to the target referent (i.e., whether the target was tall for a cup): hearers identified the target without problems in both types of situations. The time subjects took before looking at the target for the first time was measured, and although these latency times were somewhat greater when the referent were not intrinsically tall than when they were, the average difference was tiny at 554 versus 538 milliseconds. Because latency times are thought to be sensitive to most of the problems that hearers may have in processing a text, these results suggest that, for dimensional adjectives, it might be forgivable to disregard global context. This could be regarded as a (partial) vindication of Bierwisch's principle.

Testing the pragmatic constraints of section 9.2, we were able to confirm the idea that when base forms were used, the gap tends to be large. However, the Dichotomy constraint did not hold up well: even when comparing just two things, the superlative form was often preferred over the comparative [van Deemter, 2004]. Similarly, the Minimality constraint turned out to be difficult to confirm: even when the gap was large, base forms were often dispreferred. It is possible that these results are the result of quirks in the experimental setup, and that they might fail to be representative of naturalistic language use; we are inclined to treat this as an open question.

Different generation strategies could be chosen on the basis of these results, which were broadly replicated in [Barr et al., 2013] (see the corpus at `http://staff.science.uva.nl/~raquel/xprag/`). For example, one might use the superlative all the time, because this was – surprisingly – the most frequent form overall. Alternatively, one might use the base form whenever the gap is large enough, as was done in the algorithm presented in the following section.

Section 9.6 will discuss a different line of empirical investigation, but before we go there, we need to talk about algorithms.

9.4 Computational Generation of Vague Descriptions

Previous chapters have demonstrated how classic REG algorithms generate Logical Forms that individuate a referent by forming a conjunction of properties. It is important that these properties do not have borderline cases: if there was an entity x for which it is unclear whether a property P is true, then the

presence of P in a conjunction could leave it unclear whether x is part of the referent set or one of the distractors.

Although presentations of the classic algorithms frequently use examples that involve attributes such as SIZE, in fact these are treated as if they did not contain borderline cases and as if they were not context dependent: their values, such as LARGE and SMALL, always apply to the same objects, regardless of what other properties occur in the Logical Form, and without any borderline entities. This approach does not do justice to gradable adjectives, whether they are used in the base form, the superlative, or the comparative. Suppose one set a fixed quantitative threshold, making the word "large" true of everything above threshold and false of everything at or below it. Then there would be little use for this property at all. For example, suppose we are talking about dogs: then every chihuahua might be small and every alsatian large, making the combinations $\{large, chihuahua\}$ (which now denotes the empty set) and $\{large, alsatian\}$ (in which the size property is redundant) useless. Context dependency makes a property much more widely applicable.

In the next section, we show how REG algorithms such as the Incremental Algorithm (IA) can be modified to produce vague descriptions. The algorithm uses a richer type of Logical Forms than the classic algorithms, going beyond conjunctions of properties, and was implemented in a PROLOG program called VAGUE, which combines the ideas described below, with some simple rules for Linguistic Realization, which map Logical Forms to English NPs.[2]

Expressing one vague property. We start by addressing the way in which gradable information is represented in the Knowledge Base that forms the input to REG. Initially, we focus on situations in which the referent is a single entity. We assume that gradable properties are stored as attributes with decimal numerical values. We take them to be of the form n cm, where n is a positive real number, as in the example Knowledge Base below. For simplicity, we assume these numerical values to be accurate and precise.

From input of this kind, the IA is able to generate a Logical Form such as {yellow,mouse,9cm}, exploiting the attribute SIZE. The result could be the NP "The 9-cm yellow mouse", for example. The challenge, from the point of view of the present chapter, is to avoid unnecessary precision, by avoiding numerical

2 Code and documentation of the VAGUE program can be downloaded from `http://homepages.abdn.ac.uk/k.vdeemter/vague.html`.

values unless they are necessary for the individuation of the target. This challenge will be met using a *replacement* strategy. Numerical values such as 9 cm will be replaced by a superlative value that means "being the unique largest element of C". The resulting list of properties can be realized linguistically using the superlative, the comparative, or the base form ("The *largest* yellow mouse", "The *larger* yellow mouse", or "The *large* yellow mouse").

Exploiting numerical properties, singular. To ensure that Logical Forms contain a property expressible as a noun, we assume that the TYPE attribute is more highly preferred than others. Suppose also, for now, that properties related to size are less preferred than others. As a result, all the other properties that turn up in the NP – and that determine the set of objects in comparison with which an object's size will be determined – have already been selected when size comes along. Suppose the target is c_4:

TYPE(c_1) = TYPE(c_2) = TYPE(c_3) = TYPE(c_4) = mouse
TYPE(p_5) = rat
SIZE(c_1) = 6 cm
SIZE(c_2) = 10 cm
SIZE(c_3) = 12 cm
SIZE(c_4) = SIZE(p_5) = 14 cm

The first property that makes it into L is "mouse", which removes p_5 from the context set. (Result: $C = \{c_1, ..., c_4\}$.) Now SIZE is taken into account, and $size(x) = 14$ cm singles out c_4. Using the letter L for Logical Forms, we have

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

This could be the end of the matter, because the target has been singled out. But we are interested in alternative Logical Forms, to enable the generation of Vague Descriptions. One way in which such a list can be computed is as follows. Given that 14 cm is the greatest size of any mouse in the Knowledge Base, $size(x) = 14$ cm can be replaced, in L, by the property of "being the one object larger than all other elements of C" (notation: $size(x) = max_1$, where C is the set of mice). Because this is about being the largest *mouse*, rather than the largest animal, it becomes essential that L is an ordered list (rather than an unordered set), whose second property is applied to the extension of the first.

$L = \langle mouse, size(x) = max_1 \rangle$

This means: from the set of mice, pick out the largest one. Linguistic Realization will convert this into English words, for example "the largest mouse".

Exploiting numerical properties, plural. Plural references force considerable complications in the REG algorithm. To show why the computational modelling of referring can be an algorithmically more complex affair than suggested by the pseudo-code displayed in this book so far, we discuss some of these complications. Readers who wish to skip these are invited to examine Algorithm 16, then move on to section 9.5.

If plural descriptions were generated using this replacement strategy, then it would be impossible to use size to refer to sets whose elements have different sizes. To make this possible, we use inequalities, that is, values of the form $> \alpha$ or $< \alpha$, instead of values of the form $= \alpha$. Therefore, we compile the Knowledge Base into a more elaborate form by replacing equalities by *in*equalities of the form $size(x) > \alpha$ or $size(x) < \alpha$. The new Knowledge Base can be limited to *relevant* inequalities only: the ones that compare the size of an element to the sizes of other elements in the Knowledge Base:

SIZE(c_4), SIZE(p_5) $>$ 12 cm
SIZE(c_3), SIZE(c_4), SIZE(p_5) $>$ 10 cm
SIZE(c_2), SIZE(c_3), SIZE(c_4), SIZE(p_5) $>$ 6 cm,

where SIZE is an attribute, $>$ 12 cm, $>$ 10 cm, and $>$ 6 cm are values, and c_2, c_3, c_4, c_5, p_5 are domain objects of which a given $\langle Attribute, Value \rangle$ combination is true. The procedure is logically analogous to the treatment of negations and disjunctions in chapter 8: properties that are implicit in the Knowledge Base are made explicit to facilitate REG.

Given inequalities, an object may have numerous values for the same attribute; c_4, for example, has the values $>$ 6cm, $>$ 10 cm, and $>$ 12 cm. To avoid redundancy, we prefer more informative (i.e., logically stronger) inequalities over less informative ones. For example, $size(x) > 12cm$ precedes $size(x) > 10cm$ in the Preference Order, so once a size-related property is selected, later size-related properties do not remove any distractors and will therefore not be included in the Logical Form.

Suppose the target set S in our example is $\{c_3, c_4\}$. The Knowledge Base gives its two elements different sizes, hence they do not share a property of the form $size(x) = \alpha$. They do, however, share the property $size(x) > 10$ cm. This property is exploited by IA_{Plur} to construct the Logical Form

$L_1 = \langle$mouse, $>10cm\rangle$,

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

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

The case in which the numeral is 1 corresponds with the singular (e.g., "the largest mouse"). Optionally, we can go a step further to generate a truly vague description (i.e., with borderline cases). This can be done by replacing $size(x) = max_2$ by the less specified property $size(x) = max$, which abbreviates "being a set of cardinality greater than 1, all of whose elements are larger than all other elements in C". The result, which might be realized as "the largest mice", loses information, because it is no longer clear how many mice the speaker is talking about:

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

Even if comparative properties are at the bottom of the Preference Order, and more informative inequalities precede less informative ones, the order is not fixed completely. Suppose, for example, that the Knowledge Base contains information about HEIGHT as well as WIDTH, then we have inequalities of the forms HEIGHT $> x$, HEIGHT $< x$, WIDTH $> x$, and WIDTH $< x$. Which of these should come first? Greater *differences* are most likely to be chosen, presumably because they are more striking [Hermann and Deutsch, 1976]. This idea may be implemented as follows. First, the values of the different attributes should be normalized to make them comparable. Second, Preference Order should be calculated dynamically (i.e., based on the current value of C, and taking the target into account), preferring larger gaps over smaller ones. (It is possible, for example, that WIDTH is most suitable for singling out a *brown* bear, but HEIGHT for singling out a *white* bear.) The rest of the algorithm remains unchanged.

The replacement strategy is essentially a simple kind of logical inference: L_1 and L_2, for instance, are guaranteed to single out the same set, given that exactly two mice are larger than 10 cm. Given the Knowledge Base, the two lists are co-extensive. Logical inference was not an option in the classic REG algorithms, where Logical Forms were conjunctions of atomic properties, but when Logical Forms become more complex, inference becomes an important

instrument for changing the shape of a Logical Form (keeping its extension constant). We have seen examples of this in section 8.8, where avoidance of surface ambiguities was an important consideration; chapter 10.2 will show other uses of logical inference.

Expressing several vague properties. If the Knowledge Base contains several gradable attributes, a description can make use of several of them, as in (i). Even if only one gradable attribute is represented, combinations are possible, as in (ii). Let's see how REs of this kind can be generated.

(i) the flats that are taller than 8 m and less than 14 years old.
(ii) the flats whose age is between 6 and 14 years.

When opposites are part of the Knowledge Base, equalities arise automatically, as combinations of opposites: every equality of the form $height(x) = n$ metres is equivalent to the combination of a property $height(x) > i$ metres and a property $height(x) < j$ metres. Assuming that the following Knowledge Base is derived from one that contained only equalities (as explained above), it follows that every element is 6, 10, 12 or 14 years old. Consequently, if the age of an entity lies between 6 and 12 years, then its age must be 10 years.

TYPE(c_1) = TYPE(c_2) = TYPE(c_3) =TYPE(c_4) = flat
TYPE(p_5) = hall

AGE(c_1) < 10 years
AGE(c_1), AGE(c_2) < 12 years
AGE(c_1), AGE(c_2), AGE(c_3) < 14 years

AGE(c_4), AGE(p_5) > 12 years
AGE(c_3), AGE(c_4), AGE(p_5) > 10 years
AGE(c_2), AGE(c_3), AGE(c_4), AGE(p_5) > 6 years

Different measures have to be taken when *several* vague attributes are involved. Suppose, in addition to the facts represented above, the height of c_1 is 7m, the heights of p_5 and c_3 are 8m and 9m respectively, and those of c_2 and c_4 are 10m each. After recompiling these into inequalities, this yields

HEIGHT(c_1) < 8 m
HEIGHT(c_1), HEIGHT(p_5) < 9 m
HEIGHT(c_1), HEIGHT(c_3), HEIGHT(p_5) < 10 m

HEIGHT(c_2), HEIGHT(c_4) $>$ 9 m
HEIGHT(c_2), HEIGHT(c_3), HEIGHT(c_4) $>$ 8 m
HEIGHT(c_2), HEIGHT(c_3), HEIGHT(c_4), HEIGHT(p_5) $>$ 7 m

Suppose the target set is $\{c_2, c_3\}$. The algorithm starts selecting *flat*, because crisp properties are preferred over vague ones. (Result: $C = \{c_1, c_2, c_3, c_4\}$). Depending on the Preference Order, the following Logical Forms are possible:

$L_a = \langle$ flat, AGE < 14 years, HEIGHT > 8 m $\rangle$, which can be realized as (i) above ("flats that are taller than 8 m and less than 14 years old").

$L_b = \langle$ flat, AGE > 6 years, AGE < 14 years $\rangle$, which can be realized as (ii) above ("flats whose age is between 6 and 14 years").

Optionally, further modifications are possible, transforming comparative into superlative properties, for example. Once it is known which of all these outcomes yields the most felicitous RE, the algorithm may be fine-tuned. Note that if felicity is defined in terms of humanlikeness (i.e., simulating speakers) then, this time around, a *transparent* corpus (i.e., a corpus in which text is coupled with data, as was used extensively in chapter 5) may not be required; frequencies in a non-transparent text corpus (like the BNC, for example) can tell us which of these language patterns are most frequent. This is important because non-transparent corpora are far easier to obtain than transparent ones.

The model outlined in this section, and implemented in the VAGUE program, can be summarized as in Algorithm 16.

9.5 Puzzles for Incremental Content Determination

Gradable concepts cause an interesting conundrum for incremental Content Determination: if human speakers perform Content Determination incrementally, then why are properties not expressed in the same order in which they were selected? Consider a speaker referring to a bear, for example. The algorithm above suggests that the speakers starts deciding to call it a bear, to call it brown, and only after this does she decide whether it's the largest of the brown bears. This makes sense, because a gradable adjective can only be interpreted when a comparison set (e.g., the set of brown bears) is given. Yet, in English, we say "the large brown bear", rather than "the brown bear large". Gradable properties are selected last but realized first.

Algorithm 16 Generating Vague Descriptions

Input: A domain of objects encoded in a Knowledge Base that uses attributes with numerical values. The domain includes a non-empty target referent set S and one or more distractors.

Output: an NP that describes S, using gradable adjectives to express information gleaned from numerical values. Depending on the choices made in line 4, the NP can be a distinguishing description or not; likewise, the NP can contain a numeral or not.

1: Recompile the Knowledge Base, replacing equalities by inequalities, for all gradable attributes.
2: Determine the Preference Order between the different groups of attributes, for instance, giving all gradable attributes lower preference than all non-gradable ones.
3: Run an algorithm for generating references to sets (e.g., IA$_{Plur}$ of chapter 8), resulting in a list of properties that jointly identify the target set.
4: Optionally apply inferences to the list of properties. For example, replace combinations of inequalities by one exact value; replace inequalities by cardinalities.
5: Perform Linguistic Realization.

One might think that this is just how it is, because of syntactic constraints that govern the structure of English NPs. However, eye-tracking experiments cast doubt on this account, given that speakers start speaking (e.g., saying the word "large") while still scanning distractors [Pechmann, 1989]. How can speakers start uttering an adjective before it's clear that it's going to be useful?

A similar problem is discussed in the psycholinguistics of *interpretation* [Sedivy et al., 1999]: like speech production, comprehension is widely assumed to proceed incrementally, but an adjective in a vague description can only be fully interpreted when its comparison set is known. Sedivy and colleagues resolve this quandary by positing a *revision* approach to the production of vague descriptions, whereby later words allow hearers to refine their interpretation of gradable adjectives: upon hearing "large", the speaker forms a mental image of what this might mean; upon hearing "green car", this mental image is revised. Something analogous could be true for the *production* of REs. If REG algorithms are to do justice to these findings, then the proposed generation algorithm could be replaced by one that works in two phases, the second of which will sometimes revise decisions taken during the first one. Details of this idea would have to be fleshed out.

A different kind of problem is posed by multi-dimensionality: when objects are compared in terms of several dimensions, these dimensions can be weighed in different ways. Let us focus on references to an individual referent, starting

with a description that contains more than one gradable adjective. The NP "the tall fat giraffe", for example, can safely refer to an element b in this situation:

$$\begin{aligned}&\text{HEIGHT}(a) = 5 \text{ m}\\ &\text{HEIGHT}(b) = \text{HEIGHT}(c) = 15 \text{ m}\\ &\text{WIDTH}(a) = \text{WIDTH}(b) = 3 \text{ m}\\ &\text{WIDTH}(c) = 2 \text{ m}\end{aligned}$$

Cases like this would be covered by the decision-theoretic property of Pareto Optimality (e.g., [Feldman, 1980]): an object $r \in C$ is Pareto-Optimal for a given combination of Attributes if it exceeds all its competitors in at least one Attribute and none of its competitors exceeds it in any Attribute. In our example, b is Pareto-Optimal, because b is both taller and wider than its competitor, so b can be called "the tall fat giraffe". It seems likely, however, that people use doubly-graded descriptions more liberally. For example, if the example is modified by letting $\text{WIDTH}(a) = 3.1$ m, making a slightly fatter than b, then b might still be the only reasonable referent of "the tall fat giraffe". Alternative strategies are possible. The Nash Arbitration Plan, for example, would allow a doubly-graded description whenever the product of the values for the referent r exceeds that of all distractors [Nash, 1950].

We shall see in section 9.7 that multidimensionality, and the problems associated with it, can slip in through the backdoor, via the gradable notion of *salience*. But first, let's discuss another way in which multidimensionality can appear. Consider a seemingly one-dimensional word such as *big*, for example. If there existed a canonical formula for mapping three dimensions into one (e.g., length times width times height), then the result would be *one* dimension, (OVERALL-SIZE), and the algorithm discussed above could be applied *verbatim*. But it is far from clear that such a formula exists.

9.6 A Case Study: Real-World Objects and Their Sizes

Language is often studied in artificially tidy situations, for example when experiments use small scenes populated by stylized objects, presented in a small number of sizes and colours. The advantage is that the experimenter controls all the differences between the objects in the scene, so incidental features are unlikely to influence the outcome of the experiment. But this tidiness comes at a price, because real communication is seldom tidy (cf., our discussion of ecological validity in section 3.7).

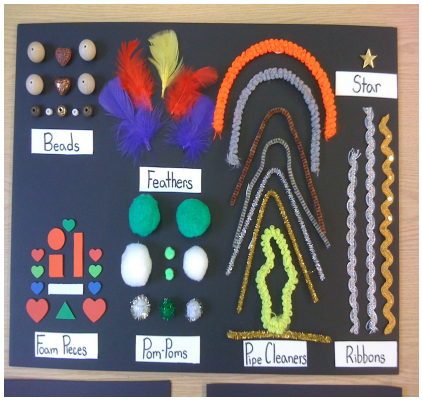


Figure 9.1
The domain that gave rise to the crafts corpus of [Mitchell et al., 2010].

To obtain an insight into real-life communication, Margaret Mitchell performed a series of experiments involving more complex objects and a playful, game-like task [Mitchell et al., 2010]. An initial elicitation study had speakers describe three-dimensional objects from an arts-and-crafts domain whose objects varied in texture, material, colour, and sheen, as well as shape and size. The domain consisted of a board on which were attached 51 objects of 7 different types (14 pieces of foam, 11 beads, 9 pom poms, 8 pipe cleaners, 5 feathers, 3 ribbons, and 1 star; see Figure 9.1). Because most objects had duplicates, speakers produced both indefinite and definite NPs ("a green pom pom", "the green pom pom"). These experiments gave rise to some new algorithms and other insights, which will be discussed in the Epilogue.

Consider the size of an entity. In the previous sections, we pretended that things have only one size. In reality, of course, they have at least three size dimensions.[3] Consequently, an entity's size may be talked about in terms of just one of these dimensions (as when we say "the long mouse"), a combination of two dimensions ("the fat mouse"), or a combination of all its dimensions ("the big mouse"). Mitchell was interested in knowing how speakers choose between different perspectives on size. Although these perspectives are not always expressed in a single adjective (e.g., speakers might say "long and

3 Even this is a simplification: a mountain doesn't have one height, but many, at different latitudes and longitudes. The domain underlying the brownies corpus – unlike the crafts corpus – avoids these complications because it is limited to simpler shapes.

rather thin", or use a relative clause, as in "which is taller than ..."), for simplicity we will continue to use the term adjective: this word will cover all these ways in which size dimensions can be conveyed in English.

One step was to elicit human-produced descriptions of objects as described above [Mitchell et al., 2010]. This led to a corpus of REs such as the longer silver ribbon, "the yellow feather", the small green heart, and so on. The **crafts** corpus, as I will call it, was used for evaluating two different algorithms, one of which was based on hand-crafted rules, the other on decision trees obtained from Machine Learning. Below we introduce these algorithms one by one. Each rests ultimately on the same data – and so of the same insights – obtained from a second corpus, which I call here the **brownies** corpus.

The elicitation experiment that produced the brownies corpus resembles earlier REG experiments. Human participants were shown pairs of rectilinear objects. Each pair consisted of two pictures of objects of the same size: two brownies, two books, two boards or two sponges, whose dimensions were carefully doctored to vary in height (the Y dimension) and width (the X dimension). The pictures made it possible to infer information about depth, but both algorithms ignore this third dimension. This second elicitation experiment gave rise to REs such as "the taller sponge", "the shorter and slightly wider board with a diagonal top side", "the smaller board", and " the most square brownie". Focussing on the adjective, there are 6 possibilities: the adjective can be looking at the 2 dimensions *overall* or (cf., [Landau and Jackendoff, 1993]) *individuating*. In the latter case, the adjective can focus on the X or the Y dimension. Because the adjective expresses a comparison, it can say that the target is either greater (positive) than the distractor or smaller (negative):

$\langle Y, positive \rangle$. For instance, *taller, longest*
$\langle Y, negative \rangle$. For instance, *shorter*
$\langle X, positive \rangle$. For instance, *thicker, wider*
$\langle X, negative \rangle$. For instance, *thinner, narrowest*
$\langle$overall$, positive \rangle$. For instance, *larger, bigger, huge*
$\langle$overall$, negative \rangle$. For instance, *tiny, smallest*

The brownies corpus, annotated to reflect the scheme above, informed both algorithms. They are discussed in the following section.

Algorithm based on hand-crafted rules. Testing on the brownies corpus confirmed three hypotheses that were formulated before the experiment. In the wording of [Mitchell et al., 2011c],

a. When two dimensions differ in the same direction between a referent object and another object of the same type, an overall size modifier will be produced more often than an individuating size modifier.
b. When two dimensions differ in opposite directions between a referent object and another object of the same type, an individuating size modifier will be produced more often than an overall size modifier.
c. The closer the aspect ratio[4] of an object, the more likely participants are to use an overall size modifier.

These findings were encoded directly into Algorithm 17. $X(r)$ is the width of the referent, $Y(d)$ the height the distractor, and so on. Having replicated Hermann and Deutsch's result, Mitchell also makes sure that if two objects differ along two dimensions in opposite directions, then the dimension representing the largest of the two differences is expressed (lines 5-6; see also section 9.4).

Algorithm 17 A part of Mitchell's algorithm for dimension choice

Input: A visual domain containing the target referent r and one distractor, d. We use $X(r)$ as short for the width of r and $Y(r)$ for the height of r (and analogously for d).
Output: A decision as to whether r is referred to using an *overall* or an *individuating* expression, and whether it is positive or negative. *largest-dim-diff* is the dimension (i.e., X or Y) with the largest difference between r and d; sign(*largest-dim-diff*) is the sign (i.e., positive or negative) of *largest-dim-diff*.

```
1: if Y(r) > Y(d) and X(r) > X(d)  then
2:     select an expression of the type [overall, positive]
3: if Y(r) < Y(d) and X(r) < X(d) then
4:     select an expression of the type [overall, negative]
5: if Y(r) > Y(d) and X(r) < X(d)  then
6:     select an expression of the type [largest-dim-diff, sign(largest-dim-diff)]
7: else
8:     ...
9: if Y(r) < Y(d) and X(r) = X(d) then
10:    ... and so on (9 cases in total) If
```

The aforementioned experiments involved only one distractor, but the algorithm was generalized by averaging the X and Y values of all distractors that have the same type as the referent; thus, in a deviation from the algorithm of section 9.4, the new algorithm sometimes calls a thing tall even though some of its distractors are taller; whether this matches human production is a question that has yet to be settled.

4 Aspect ratio is defined as x/y, where x is the value of the width and y of the height dimension.

Algorithm from Machine Learning. We have seen before (e.g., section 6.6) that when transparent corpora like the crafts and brownies corpora are available, they can be exploited by a Machine Learning program that looks for regularities in the corpus. One way to do this is to learn *decision trees*, which use boolean combinations of data features (with suitable values for these features) to predict when a given type of RE can be used; tree structures are employed to make sure that the most informative features come first. [Mitchell et al., 2011b] learned decision tree classifiers from the brownies corpus. Mitchell decided to use not only the most obvious factors (such as $X(r)$, $Y(r)$, $X(d)$ and $Y(r)$, which contain the dimensions of the target and the distractor), but also some "smart" ones that had emerged from studying the crafts domain. Examples include *target height divided by target width*, *target height minus target width*, and the following ones. The value is always a number; as in the extended algorithm (a), $X(d)$ and $Y(r)$ are averaged over all distractors.

xratio = target width / distractor width
yratio = target height / distractor height
discx = 1 if $X(r) > X(d)$; 2 if $X(r) = X(d)$; 3 if $X(r) < X(d)$
discy = 1 if $Y(r) > Y(r)$; 2 if $Y(r) = Y(d)$; 3 if $Y(r) < Y(d)$

An example of a classifier that was learned is the following tree for the use of type [overall, negative] (i.e., adjectives like *small*):

```
discy =1: no
discy >1
  discx <=1: no
  discx >1
     drat <= 1
        xratio <= 0.909: yes
        xratio >0.909
           discy <=2: no
           discy >2
              ... (etc.)
```

This classifier is typical for using "smart" features in prominent positions near the top of the tree, so whatever results are obtained by the Machine Learning is heavily indebted to the hypothesis-testing that has informed the other algorithm as well.

Both algorithms were evaluated on both the brownies and the crafts corpus, using an approach that resembles the Dice-based one in chapter 5. The

most meaningful evaluation is the one that compares with the crafts corpus, because it is different from the one on which the classifiers were trained. [Mitchell et al., 2011c] reports both precision and recall (which the Dice metric combines) but because the REs in the corpora only rarely contained more than one size adjective (i.e., one rarely finds REs like "the feather that's *long* and *thin*"), precision and recall were always nearly identical. The evaluation results were good, with precision and recall at about 81% for both algorithms. By comparison, an *oracle* that always chooses, for a given referent, the type of adjective (e.g., [overall, positive]) that was chosen most frequently by the speakers of the crafts corpus when referring to that referent, scores 89.1%. This oracle would be very difficult to beat of course. In other words, the two algorithms were both successful, and to almost exactly the same degree.

Mitchell's approach looks at REG from a very different perspective compared to section 9.4. Consider one of the 5 feathers in the crafts domain. Suppose its two dimensions are in agreement with each other: the referent is both the second-longest and second widest feather, with the same distractor being both longer and wider. Mitchell's algorithms will tend to choose [overall, positive]. This does not mean, however, that the referent must necessarily be called "the *large* (*big*, *sizeable*) feather"; instead, it might be called "the *second* largest feather", or "this rather large feather". Unlike the VAGUE program, these algorithms do not go as far as generating the semantic content of a complete RE: they only decide which of the 6 attributes mentioned above will be used: [individuating, positive], [individuating, negative], and so on. No actual words are chosen. More importantly, it has not been decided where in the relevant dimension the referent is to be placed. Finally, these algorithms do not tell us whether a given attribute identifies the referent. To do that, an algorithm like the one of section 9.4 needs to be added. Perhaps the best way to see these algorithms, therefore, is as a way to decide what kind of size attribute to consider when looking for a suitable combination of properties (e.g., what size attributes to store in the Preference Order of the Incremental Algorithm).

A question that neither of these approaches addresses in earnest is what differences are large enough to matter.[5] The (a) algorithm, for instance, contains a rule of the form "If $Y(r) < Y(d)$ and $X(r) = X(d)$, then ...". This rule fires if two values are equal, but it is unclear whether a tiny difference of, say, 1 mm matters enough to stop them from counting as equal.

5 This is related to the notion of a Just-Noticeable Difference (JND) in perception research, e.g., [van Deemter, 2010], chapter 8.

We have focussed on size, but other attributes are multidimensional as well: suppose we want to refer to a person, for instance, focussing on her quality as a researcher. We can call her productive or creative or smart, for instance, or we can say that her work has had much impact. Or, instead of focussing on such individual dimensions, we can choose an "overall" adjective, calling her a good researcher, for instance. The choice between all these perspectives resembles the choice of perspective associated with the size of an object.

Furthermore, we have focussed on REs, but the choices on which this section has focussed are equally relevant for indefinite noun phrases ("a tall building") and quantified noun phrases ("every tall building"), whatever role they play in a sentence. We hypothesize that these choices depend on the same factors and can be made in similar ways.

9.7 Can We Ever Be Clear? Salience as a Gradable Property

We have seen in section 4.9 that salience is best regarded as a gradable property, which some things possess to a greater degree than others. We can now see that a natural treatment of salience falls automatically out of the treatment of vague descriptions presented in section 9.4. This insight will allow us to gain a more uniform understanding of REG. However, we shall also see that the gradability of salience makes almost all reference error prone.

Essentially, Krahmer and Theune's proposal (see section 4.9) analysed "the black mouse" as denoting the unique *most salient* entity in the domain that is both black and a mouse. Now suppose we let REG treat salience just like other gradable attributes. Suppose there are ten mice, five of which are black, whose degrees of salience are 1, 1, 3, 4, and 5 (the last one being most salient), whereas the other objects (cats, white mice, etc.) have a higher salience. Then the algorithm of section 9.4 might generate this list of properties:

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

This is a distinguishing description of the black mouse whose salience is 5: "the most salient black mouse". The simpler description "the black mouse" can be derived by stipulating that the property of being most salient is normally left implicit in English as well as, presumably, in other languages.

It is now easy to see why plural descriptions are often ambiguous. Taking salience into account as suggested above, the singular "the black mouse" can only refer to the most salient mouse. But "the mice" can refer to the most

salient two (sized 5 and 4), the most salient three (sized 5, 4 and 3), or to all of them. To disambiguate the description, a number can be used (e.g., "the two mice"), just like in the case of vague descriptions.

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

TYPE: *d1* (dog), *d2* (dog), *d3* (dog), *d4* (dog), *c5* (cat)
SIZE: *d1* (20 cm), *d2* (50 cm), *d3* (70 cm), *d4* (60 cm), *c5* (50 cm)
SALIENCE: *d1* (6), *d2* (4), *d3* (3), *d4* (5), *c5* (6).

If we are interested in the *three* most salient dogs (d_1, d_2, and d_4), then "the large(est) dog" designates *d4*, but if we are interested in the *four* most salient ones (d_1, d_2, d_3, and d_4), then it designates d_3. In other words, the description is ambiguous between *d3* and *d4*, depending on whether we attach greater importance to salience or size. This is borne out by our generation algorithm. Consider the simpler of the two treatments of salience, for example, which starts out with a reduced domain. If d_4 is the target, then the reduced domain (consisting of all things at least as salient as the target) is $\{d_1, d_2, d_4, c_5\}$; "dog" narrows this down to $\{d_1, d_2, d_4\}$, after which size $= max_1$ generates "the large dog". But if d_3 is the target, then the same procedure applies, starting with the full domain (because no element is less salient that d_3) and the same description is generated to refer to a different animal. For readers, csalience and gradable adjectives are a problematic combination.

But this is not all, because salience itself is multidimensional (see e.g., Paul Piwek's [Piwek, 2009] for a 3-dimensional account). Consider a situation in which two people agree to meet in a coffee bar (Figure 9.2); they are near a bar that is so decrepit that it is barely functional; much further away (making it less salient) is another coffee bar that is large and attractive (making it more salient). In such a situation, it can be unclear which of the two coffee places is intended. We usually assume that the entities around us can be identified by suitable REs, but once salience is taken into account (especially in combination with plurals and/or other gradable dimensions) it becomes difficult to generate descriptions that are immune to being misunderstood. It might be thought that referential ambiguity can be avoided by making the degree of salience of all domain objects explicit, for example, by saying "let's pay attention to buildings in area X, and to no other buildings". However, this begs the problem, because the RE "area X" itself is potentially subject to referential ambiguity.

Figure 9.2
Multi-dimensional salience as a source of ambiguity: "Let's meet at the coffee bar."

In other words, despite computer scientists' neat abstractions – involving domains, distractors, and distinguishing descriptions – communication is often risky; we will see this again in chapter 14.

9.8 Summary of the Chapter

Until this chapter, numbers were absent from the information from which referring expressions were generated: the shared Knowledge Bases that form the basis of the classic algorithms, for example, did not contain numerical values. By examining the effect of numerical values on reference, the present chapter has opened the door to a large area of research, with potential relevance for "big data" applications, where numerical information is paramount.

- The semantics of vague adjectives in REs can be likened to that of superlatives. [Section 9.1] There are, however, pragmatic differences between them, which we have dubbed the principles of Small Gaps, Dichotomy, and Minimality. [Section 9.2]

- The monotonic approach to REG that dominates earlier chapters of this book can be extended to cover gradable attributes. The resulting algorithms, however, tend to be considerably more complicated than the classic REG algorithms. [Section 9.4]
- The algorithm discussed in this chapter should not be seen as a model of the human production *process* (see section 16.1, where process models are contrasted with product models). Experiments by Sedivy et al. suggest that, to model human production, a more complex process may be needed, which generates gradable adjectives twice: once as a first "draft" and once for real. [Section 9.4]
- Size is multi-dimensional. Consequently, speakers can express a single size dimension or a combination of size dimensions (e.g., as expressed by the adjective "big"). A case study by Margaret Mitchell gave insight into some of the principles underlying this choice. It would be interesting to see whether these principles generalize to other multi-dimensional attributes, like the health of a person or the quality of a book. [Section 9.6]
- Three principles appear to govern the choice between individuating and overall size descriptions: (a) When two dimensions differ in the same direction, then *overall* size modifiers are more probable than *individuating* ones. (b) When two dimensions differ in opposite directions, then *individuating* size modifiers are more probable than *overall* ones (c) The closer the aspect ratio of an object, the more probable are *overall* size modifiers as compared to *individual* ones. [Section 9.6]
- Confirmation was found for Hermann and Deutsch's principle that if two objects differ along two dimensions in opposite directions (e.g., the referent is taller but thinner than the distractor), then the dimension that represents the largest of the two differences tends to be expressed. [Section 9.6]
- Combinations of gradable properties are difficult to interpret. Salience is gradable, so this makes combinations of *one* gradable adjective with salience potentially unclear. Salience is multi-dimensional, causing it to behave as a combination of gradable attributes, so this makes the use of all REs potentially unclear (as in "the coffee bar", said when one coffee bar is nearer whereas another one is larger). [Section 9.7]
- These findings suggest that we may have to give up the illusion that referring expressions must be unambiguous descriptions in all situations. This will be an important theme in Part IV of this book.

10 Fourth Extension: Exploiting Modern Knowledge Representation

We have seen how recent work has started to go beyond the classic REG problem, by generating expressions that make use of relational and gradable properties, referring to sets, and using proper names. Earlier on (especially in chapter 6), we saw how a variety of computational approaches have been applied to REG, such as Labelled Directed Graphs, Constraint Satisfaction, and Bayesian Reasoning; we have also seen various uses of Machine Learning. All these developments have made REG algorithms more interesting and useful models of reference production.

But despite these advances, there are still many situations in which a human speaker would be able to utter a distinguishing description, yet none of the algorithms discussed so far is able to do this. In this chapter and the following, we shall ask how REG might become *logically complete*. The key to the extensions discussed in this chapter lies in techniques that have been developed over the last decades in Knowledge Representation (KR) and that have come to be associated with formal ontologies and their practical applications. I consider it unfortunate that KR and Formal Logic have come to be sidelined in much of Computational Linguistics (cf., the start of chapter 5), in line with the "shrinking horizons" that Ehud Reiter warned against [Reiter, 2007]. What follows is an attempt to redress the balance, focussing on reference, but in a manner that can have repercussions for phenomena other than reference.

By using modern KR techniques, we shall construct richer Knowledge Bases and generate richer Logical Forms than before, which will allow our algorithms to refer in situations where earlier algorithms were not able to do this. Furthermore, KR will help REG to exploit ontological information. Ultimately, it will also become possible to refer to entities whose existence is not stated explicitly in the Knowledge Base, but only implied, allowing these algorithms to generate the *attributive* descriptions noted by philosophers and other theoreticians of language (section 2.6).

The problems discussed in this chapter have not been studied for long yet, and the material presented here is more tentative than that of previous chapters. We shall focus on the expressive power of generation algorithms. This

means that logical concerns will be at the centre of our discussion. The question of what RE would be most natural, or most effective – which concerned us extensively in previous chapters – will take a backseat.

This chapter is constructed as follows[1]: First we introduce the idea of Knowledge Representation and Description Logic (section 10.2), after which we explain how Description Logic can be applied to REG, allowing a larger range of REs to be generated than before (sections 10.3, 10.4, 10.5). We then turn to a discussion of how REs may be generated that refer to entities whose existence is not directly stated but only deduced (section 10.6). Finally, looking back on the body of this chapter, we discuss why logically complex REs were, and still are, worth studying (section 10.7).

10.1 Knowledge Representation and REG

The algorithms discussed so far start from simple, tailor-made Knowledge Bases, which contain atomic information. We have seen in section 6.4 how 2-place relations can be taken into account in the generation of relational descriptions. Yet relational descriptions cry out for a further increase of expressive power. It is nice to be able to generate "the cup on the table". But why stop there, instead of pressing on to generate "the table with *two* cups on it", and "the table that has *only* one cup on it" as well?

Moreover, the information in the Knowledge Base, which represents all the knowledge that is shared between the speaker and the hearer, can be much more complex than the atomic statements (and their negations) that we have so far allowed. Some early REG algorithms made use of non-atomic information, but this never went beyond representing the fact that one 1-place property (e.g., being a dog) subsumes another (e.g., being a poodle). Here are some of the things that these Knowledge Bases are unable to express, and that the algorithms discussed so far cannot make use of when referring:

a. This painting is Flemish or Dutch.
b. The relation "part of" is transitive.
c. For all x, y, x is to the left of y if and only if y is to the right of x.
d. For all x, y, z, if x buys y and z is a part of y, then x buys z.

1 The discussion of the GROWL algorithm and related issues in sections 10.3, 10.4, and 10.5 is an elaboration of the proposal in [Ren et al., 2010], with special thanks to Yuan Ren for advice.

The disjunction (a) is what an art lover may get from inspecting a 14th-century church panel. This "incomplete" information about the painting cannot be represented if only atomic information is available (unless "Flemish or Dutch" is treated as an atomic property).

Sentences (b), (c), and (d) might be implicitly present in a Knowledge Base of atomic facts, but there is substantial mileage in representing these rules explicitly, because it allows information to be represented much more economically and insightfully. Consider (b), for example, when a complex machine is described. Suppose part p_1 is part of p_2, which is part of p_3, which is part of p_4. Axiom (b) allows us to leave it at that: four other facts (that p_1 is also part of p_3 and p_4, and that p_2 is part of p_4) can be deduced from the three facts that are represented. In large domains the savings can be enormous. Existing REG algorithms lack the formalism for expressing knowledge of this kind.

Luckily, the Knowledge Representation community has over 30 years of experience designing formalisms and algorithms that handle (some) logically structured knowledge efficiently. Not only are these able to represent many different kinds of information (including (a–d) above), they are also able to perform automated reasoning with them, for example to check whether a Knowledge Base is logically consistent. Key reasoning tasks of early systems in this tradition, such as KL-ONE [Brachman and Schmolze, 1985] were soon shown to be undecidable, but later systems are decidable, and considerable care is invested in making sure that reasoning is performed as fast as possible.

Formalisms for Knowledge Representation come in different flavours, of which Conceptual Graphs and Description Logic are probably the most important. Description Logic, for example, is now implicated in a large amount of work on formal ontologies for practical applications, such as the SNOMED ontology for medical information [Benson, 2012], and for the Semantic Web. All these formalisms are closely aligned with Predicate Logic, and Modal Logic [Baader et al., 2003]. Some of their terminology deviates from standard logical practice as a result of their evolution from different ways of thinking about knowledge that had established themselves before the link with Formal Logic was properly understood, but in most cases, the translation between formalisms is straightfoward. (For example, where Predicate Logic speaks of a 2-place relation, Description Logic speaks of a role.)

To let REG benefit from these developments, we proposed to analyse REG as a *projection* problem in Conceptual Graphs [Croitoru and van Deemter, 2007]. Independently, [Areces et al., 2008] analysed REG as a problem in Description Logic (DL). It is this latter idea, which is traceable at least to

[Gardent and Striegnitz, 2007], that we shall use as our starting point in this chapter. The idea is to generate a DL formula, called a concept expression or simply a concept, such as

$$Dog \sqcap \exists love.Cat$$

This concept denotes the set $\{x \epsilon Dog : \exists y \epsilon Cat(love(x, y)\}$, that is, the set of dogs intersected with the set of objects that love at least one cat. To apply this to REG, the idea is to check how many individuals turn out to be elements of this set (i.e., instances of the concept); if the number is 1, then the concept refers to this individual; if the number is greater, then it refers to a set. Figure 10.1 depicts a small Knowledge Base involving relations between several people and animals. Here the above-mentioned concept identifies $d1$ as "the dog that loves a cat", singling out $d1$ from the five other objects in the domain. Of

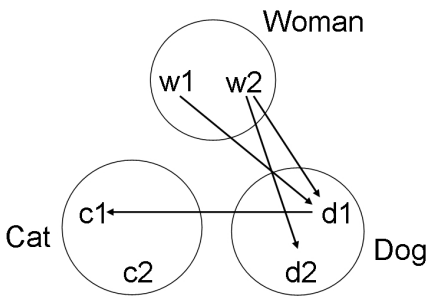


Figure 10.1

Diagram depicting a small domain. Edges from people to animals denote the relation "feed"; edges between animals denote the relation "love".

course, earlier approaches to the generation of relational REs (section 6.4) were already able to refer to $d1$, but we shall see in section 10.4 that DL can allow REG to generate a much wider range of REs.

10.2 Description Logic: a Primer

A Knowledge Base based on Description Logic [Baader et al., 2003] describes concepts (denoting sets of domain objects) and binary relations (denoting sets of pairs of domain objects). Each Description Logic embodies a compromise between expressive power and speed of reasoning; a particularly expressive Description Logic is $\mathcal{SROIQ}$ [Horrocks et al., 2006], the logic that underlies

the language OWL2, which is often used in connection with the Semantic Web. Principally, a $\mathcal{SROIQ}$ ontology consists of a Terminology Box (TBox) $\mathcal{T}$ and an Assertion Box (ABox) $\mathcal{A}$. The ABox contains axioms about specific individuals, for example,

$[a : C]$ means that a is an instance of the concept C (a has the property C).
$[(a, b) : R]$ means that a stands in the R relation to b.

The TBox contains generic information concerning concepts and relations, such as the fact that R is symmetric, irreflexive, or transitive. Most importantly, $\mathcal{T}$ can say that $C \sqsubseteq D$, where C and D are concepts, meaning that every instance of C must be an instance of D too. For example, if D is defined as the concept Person and C as the concept "Man below 15 years of age", then if Man $\sqsubseteq$ Person, it follows that $C \sqsubseteq D$. In other words, $C \sqsubseteq D$ means that our combined information about C and D allows us to conclude that every instance of C must be an instance of D.

The notion of a concept is recursively defined. First of all, atomic concepts are concepts. Furthermore, if C and D are concepts, and R is a binary relation (i.e., a "role"), then so are each of the following:

$$\top \mid \bot \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \exists R.C \mid$$
$$\geq nR.C \mid \ \leq nR.C \mid \forall R.C \mid \exists R.Self \mid \{a_1, \ldots, a_n\}$$

where n is a non-negative integer, a_i are individual constants, and R is a relation, which can be atomic or the inverse of a relation S (i.e., R may be S^-).

As one can see, negation can apply to a concept, but it it cannot apply to a relation: in $\mathcal{SROIQ}$ one cannot compose a concept $\exists \neg feed.Dog$, for example. In a Closed World, however, we can interpret $\neg feed$ as the set of all pairs of domain elements that do *not* stand in the relation $feed$. The proposal in section 10.4 will follow [Ren et al., 2010] in using an extension of $\mathcal{SROIQ}$, which we shall call $\mathcal{SROIQ}^+$. This extended language will allow negated relations as part of concept expressions, but their handling is computationally less straightforward than that of other constructs.

$\top$ ("top") is the most general concept, denoting the domain, $\bot$ ("bottom") is the least general concept, denoting the empty set. Quantifiers do not use variables. For example, $\exists R.C$ denotes the set of things that stand in the relation denoted by R to at least one element of the set denoted by C. Numerical quantifiers of the form $\geq nR.C$ extend this idea to containing at least n elements of the set; $\leq nR.C$ denotes the set of entities that stand in the relation denoted by R to at most n elements of the set denoted by C. $\forall$ may be thought of as the

English quantifier "only": $\forall R.C$ denotes the set of those x such that, for all y, if x stands in the relation denoted by R to y, then y must be an element of the set denoted by C. Finally, $\exists R.Self$ denotes the set of objects standing in the relation R to themselves; this is known as self-restriction.

We use CN, RN, and IN to denote the set of atomic concepts, atomic relations, and individuals. As usual in logic, the meaning of an expression is defined by means of an *interpretation* function. In DL it is customary to apply the interpretation function $\mathcal{I}$ to the domain of discourse as well as to the constructs of the language. Thus, $\mathcal{I}$ is a pair $\langle \Delta^{\mathcal{I}}, .^{\mathcal{I}} \rangle$ where $\Delta^{\mathcal{I}}$ is the (non-empty) domain of discourse; $.^{\mathcal{I}}$ maps each atomic concept to a subset of $\Delta^{\mathcal{I}}$, each atomic role to a subset of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, and each individual to an element of $\Delta^{\mathcal{I}}$. Thus, an interpretation is closely aligned to what is known in other areas of logic as a *model*. The interpretation of complex concepts is defined recursively, for example, $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$.

$\mathcal{I}$ is a *model of* the Knowledge Base Σ, written $\mathcal{I} \models \Sigma$, if all the axioms in Σ are satisfied in $\mathcal{I}$. Notably, Σ can have multiple models. For example if $\mathcal{T} = \emptyset, \mathcal{A} = \{[a : A \sqcup B]\}$, then there is an interpretation $\mathcal{I}_1$ such that $A^{\mathcal{I}_1} = \{a\}, B^{\mathcal{I}_1} = \emptyset$, and an $\mathcal{I}_2$ s.t. $B^{\mathcal{I}_2} = \{a\}, A^{\mathcal{I}_2} = \emptyset$. The effect of axioms such as $[a : A \sqcup B]$ on REG (which express a kind of incomplete knowledge) will be discussed in section 10.6.

To see how non-trivial inferences can be deduced from a simple Knowledge Base, consider Figure 10.1 again, assuming that the domain does not contain other individuals than the ones depicted, nor do individuals have other properties and relations than the ones depicted. From this it follows that $\exists love.Cat \sqsubseteq \geq 2\ feed^{-}.Woman$ (everything that loves a cat is fed by at least two women) and $\geq 2 feed^{-}.Woman \sqsubseteq \exists love.Cat$ (everything that is fed by at least two women loves a cat). The two concepts have the same extension. If another axiom were to assert that the two concepts are disjoint with each other, then it would follow that neither is satisfiable.

10.3 Applying Description Logic to Familiar REG Problems

In REG, as we shall see, Description Logics can be used for the dual purpose of constructing Knowledge Bases and constructing Logical Forms for REs. Description Logic forces one to be explicit about some assumptions that are usually left implicit in REG. In discussing these matters, it will be useful to

start from very simple examples, because these will permit us to see the logical issues most clearly. Consider once again the following Knowledge Base:

TYPE: dog $\{a, b, c, d, e\}$, poodle $\{a, b\}$
COLOUR: black $\{a, c\}$, white $\{b, e\}$

Description Logic is most often employed in situations where information that is not stated explicitly may still be true; there can even exist individuals that are not mentioned in the Knowledge Base. This is known as the Open World assumption. Most work on REG, by contrast, tacitly assumes that the shared Knowledge Base is a Closed World. In the Knowledge Base above, for example, if there were poodles other than a and b, then these might include c, so "black poodle" no longer identifies a uniquely.

The Closed World Assumption assumes that all atomic facts whose truth does not follow from the axioms in the Knowledge Base are false. (For example, in the example above, c is not a poodle.) In section 10.4, we shall use a more fine-grained approach that allows us to close an ontology only partly, namely, by using a DBox [Seylan et al., 2009]. A DBox $\mathcal{D}$ contains only atomic formulas. Every concept or relation appearing in the DBox is closed, and the DBox defines their extensions exactly, so if $\mathcal{D} \not\models [a : A]$, then $[a : \neg A]$. Crucially however, concepts and relations not appearing in $\mathcal{D}$ remain open, and this is where generic rules can make themselves felt, because they can make information inferable. For example, given $\mathcal{T} = \{Dog \sqsubseteq \exists feed^{-}.Woman\}$ (*every dog is fed by some woman*) and $\mathcal{A} = \{[d1 : Dog], [w1 : Woman]\}$, there must be some woman who feeds $d1$. If the domain is closed using $\mathcal{D} = \mathcal{A}$, then we can infer that this is $w1$, because this is the only woman in the domain.

If a Closed World is assumed, then given a Knowledge Base, every Description Logic concept denotes a set. If this set is not empty, then the concept can be seen as referring to this set. It is this idea [Gardent and Striegnitz, 2007] that Areces and colleagues explored, focussing on an ABox without a TBox. Using a DBox, the domain in Figure 10.1 can be formalized as follows:

$\mathcal{T}_1 = \emptyset$

$\mathcal{A}_1 = \{[w1 : Woman], [w2 : Woman], [d1 : Dog], [d2 : Dog],$
$[c1 : Cat], [c2 : Cat], [(w1, d1) : feed], [(w2, d1) : feed],$
$[(w2, d2) : feed], [(d1, c1) : love]\}$

$\mathcal{D}_1 = \mathcal{A}_1$

The algorithm proposed by Areces and colleagues computes all the *similarity sets* associated with a given Knowledge Base and a given Description Logic [Areces et al., 2008]. Similarity sets can be seen as generalizing the satellite sets of [van Deemter and Halldórsson, 2001] by taking relations into account, and by being explicitly parameterized on a given logic. The parametrization to a logic is important, because which REs are expressible, and which objects are referable, depends on the logic that provides the REs. Here we use the word "satellite" instead of the word "similarity" because this is the term we used in chapter 8, and because it rightly suggests a non-symmetrical relation: a can be a satellite of b without b being a satellite of a. Thus, if L is a Description Logic and $\mathcal{I}$ an interpretation, and i, j are elements of $\Delta^{\mathcal{I}}$, then we shall say that

> j is an *L-satellite* of i given $\mathcal{I}$ if and only if
> for every concept φ of L: if $i \epsilon \varphi^{\mathcal{I}}$, then $j \epsilon \varphi^{\mathcal{I}}$.
>
> The *L-satellite set* given $\mathcal{I}$ of an individual i is
> the set of all L-satellites of i given $\mathcal{I}$.

Clearly, the logic L contains an RE for i given $\mathcal{I}$ if and only if the L-satellite set of i given $\mathcal{I}$ equals the singleton set $\{i\}$. In other words, an RE for a given object i exists, given a Description Logic L and a model, if and only if there is no other object j in the model such that j is in the extension of every concept in L that has i in its extension.

Areces and colleagues applied their approach to two different types of Description Logic. Here we focus on the most expressive of the two, $\mathcal{ALC}$, which still permits a far smaller range of constructs than $\mathcal{SROIQ}$, excluding the $\forall$ quantifier and numerical quantifiers, for example:

$$\top \mid \neg C \mid C \sqcap D \mid \exists R.C$$

Unlike the algorithms discussed so far, theirs (Algorithm 18) does not aim to find an RE for one particular "intended" referent: it finds REs referring to *any* subsets of the domain that can be referred to given the language $\mathcal{ALC}$. The algorithm adapts an algorithm designed by Hopcroft [Hopcroft, 1971]. It finds out what sets of objects are describable through increasingly complex intersections of (possibly negated) atomic concepts, then tries to extend these intersections with concepts of the form $(\neg)\exists R1.Concept$, then with concepts of the form $(\neg)\exists R2.(Concept \sqcap (\neg)\exists R1.Concept)$, and so on. At each step, more complex concepts are constructed from concepts constructed earlier.
The algorithm calls a function Add, which checks, for each of the newly constructed concepts, whether it makes a useful addition to the set RE:

Algorithm 18 Computation of $\mathcal{ALC}$-satellite sets

Input: An interpretation function $\mathcal{I}$ that assigns denotations to elements of AtConcepts and AtRels.
Output: A set RE of formulas φ such that $\{\varphi^{\mathcal{I}} : \varphi \in \text{RE}\}$ is the set of things that are $\mathcal{ALC}$-satellite sets (of any object in $\Delta^{\mathcal{I}}$) given $\mathcal{I}$.

```
1: RE:= {⊤}
2: for each p ∈ AtConcepts do
3:     Add(p,RE)
4: while There exist φ ∈ RE such that ||φ^I|| > 1 do
5:     for each φ ∈ RE and R ∈ AtRels do
6:         Add(∃R.φ,RE)
7:     if made no changes to RE then
8:         Exit
```

Add(φ,RE)

```
1: for each ψ ∈ RE with ||ψ^I|| > 1 do
2:     if (ψ ⊓ φ)^I ≠ ∅ and (ψ ⊓ ¬φ)^I ≠ ∅ then
3:         add formulas ψ ⊓ φ and ψ ⊓ ¬φ to RE
4:         remove ψ from RE
```

Consider the Knowledge Base above (Figure 10.1), for instance. The algorithm concludes in 3 phases, resulting in increasingly fine-grained partitions of the domain, where each step builds on the previous ones. In step (3), for example, $w2$ can be identified because $d2$ is identified during step (2).

1. $Dog = \{d1, d2\}$, $Woman = \{w1, w2\}$, $Cat = \{c1, c2\}$.
2. $Dog \sqcap \exists love.Cat = \{d1\}$, $Dog \sqcap \neg\exists love.Cat = \{d2\}$.
3. $Woman \sqcap \exists feed.(Dog \sqcap \neg\exists love.Cat = \{w2\}$,
   $Woman \sqcap \neg\exists feed.(Dog \sqcap \neg\exists love.Cat = \{w1\}$.

On termination, the algorithm has partitioned the domain into minimal subsets, each of which gets coupled with an $\mathcal{ALC}$ formula denoting it. The $\mathcal{ALC}$-satellite sets that exist relative to this model are all found by the algorithm, namely, $\{d1, d2\}$, $\{w1, w2\}$, $\{c1, c2\}$, $\{d1\}$, $\{d2\}$, $\{w2\}$, $\{w1\}$. The first two of these were removed by step 4 of *Add*. Four of the animals can be individuated; the cats cannot, because inverse relations (e.g., being fed) are not considered by this version of the algorithm.

The algorithm was implemented and evaluated on the Filing Cabinet corpus of [Viethen and Dale, 2006b], with encouraging results. Being very small,

the corpus contained only 15 relational REs; transparent corpora in which relational descriptions play a significant role were sparse, which is why a small corpus had to suffice. The program managed to generate 10 of these 15 REs correctly, most of which were of the pattern "the (orange, blue,..) drawer above/-below/next to the (orange, blue,..) drawer". A typical example of an RE in the corpus that the program was unable to generate is "the orange drawer below the *two* yellow drawers". This hints at the need for REs that contain numerical quantifiers (e.g., "two"), an issue to which we shall turn presently.

The paper by Areces and colleagues is not only of interest because of this algorithm but also because it shows how REG algorithms in the literature are related to each other, by showing how their behaviour can be mimicked by applying the approach described above to different Description Logic languages. We shall not discuss these matters here any further. Instead, we shall demonstrate how the work of these authors has inspired our own proposal [Ren et al., 2010].

10.4 Exploiting the Full Power of DL

These ideas can be generalized in two ways: by using TBox reasoning, and by using a larger range of quantifiers. The two innovations are independent of each other but explained here by means of one and the same example.

Areces and colleagues took logical *models*, not axioms, as their starting point, which is equivalent to having a Knowledge Base that only contains an ABox and that is interpreted under a Closed World Assumption. Consequently, their algorithm essentially uses model-checking, rather than full DL reasoning. It is possible, however, to generalize their ideas. Suppose we extend Figure 10.1 with a rule saying that *if one is feeding an animal, and if this animal loves another animal, then one feeds the second animal too*, while also adding an edge to the "love" relation (Figure 10.2) between $d2$ and $c2$: Suppose we close the domain, using a DBox, as follows, using relation composition ($\circ$):

$\mathcal{T}_2 = \{feed \circ love \sqsubseteq feed\}$

$\mathcal{A}_2 = \{[w1 : Woman], [w2 : Woman], [d1 : Dog], [d2 : Dog],$
$[c1 : Cat], [c2 : Cat], [(w1, d1) : feed], [(w2, d1) : feed],$
$[(w2, d2) : feed], [(d1, c1) : love], [(d2, c2) : love]\}$

$\mathcal{D}_2 = \{[w1 : Woman], [w2 : Woman], [d1 : Dog], [d2 : Dog],$
$[c1 : Cat], [c2 : Cat]\}$

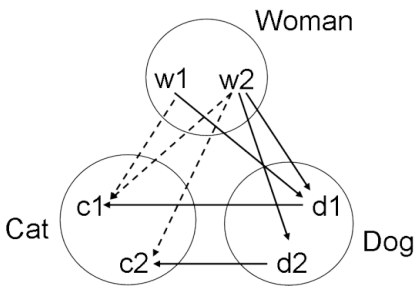


Figure 10.2

An extension of Figure 10.1. Dashed edges denote implicit relations, inferred using the TBox. As before, arrows between dogs and cats denote the relation "love"; other arrows denote the relation "feed". Note that, this time around, d2 loves c2.

The implicit facts $[(w1, c2) : feed]$, $[(w2, c1) : feed]$, and $[(w2, c2) : feed]$ can now be inferred automatically by exploiting the TBox axiom. The use of rules and reasoning is crucial from the point of view of DL: without it, DL could hardly claim to be a logic (or a piece of Knowledge Representation, for that matter): it would little more than a database formalism.

The second innovation is of greater importance from the point of view of reference modelling (i.e., REG). Note that all the REG algorithms discussed so far are very limited in terms of the relational REs they generate. The algorithm based on $\mathcal{ALC}$, for example, allows us to say "the woman who feeds a cat" ($Woman \sqcap \exists feed{\cdot}Cat$) but not "the woman who feeds two cats", allowing existential quantification but excluding other quantifiers.

If only existential quantifiers are used, then some referents cannot be distinguished at all. In fact, the algorithm fails to identify *any* individual in Figure 10.2 (also depicted in Figure 10.3), because none of their satellite sets is a singleton. For example, the satellite set of $w1$ is $\{w1, w2\}$, which has two elements. I will show presently how all the individuals in the Knowledge Base become referable if other quantifiers and inverse relations are allowed.

But first we need to ask what level of expressivity should be achieved. In attempting to answer it we can benefit from the conceptual apparatus developed in an area of (initially) Formal Logic and (later) the formal semantics of natural language, known as the theory of Generalized Quantifiers. Research on Generalized Quantifiers goes back to Mostowski and was further developed by Barwise and Cooper and by Van Benthem [Mostowski, 1957] [Barwise and Cooper, 1981] [van Benthem, 1986]. This

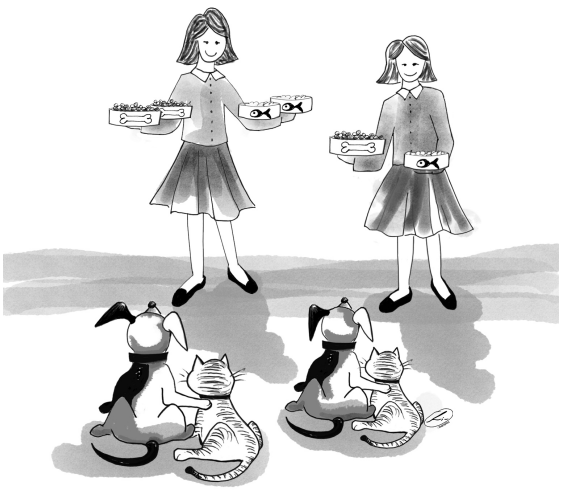


Figure 10.3
The woman who feeds two dogs is depicted to the left of the one who feeds only one.

research program seeks to understand what quantifiers are possible, in logic and in natural language, over and above the ones that happen to be best known (such as *all* and *some*). In the process, this research program has generated many useful insights and technical innovations [Peters and Westerstahl, 2006].

Essentially, a Generalized Quantifier is a relation between two sets. A Generalized Quantifier Q can occur in many different contexts, for example, in the context Q N V (where the noun N and the verb V both denote sets), as in "All (some, ten, most, etc.) cats slept". The most general format for REs that involves a relation R is "The N_1 who R Q N_2's", where N_1 and N_2 denote sets, R denotes a relation, and Q a Generalized Quantifier, as in "the women who feed *some* dogs". An expression of this form refers to an individual entity if it denotes a singleton set. Using a set-theoretic notation, this means that the following set has a cardinality of 1:

$$\{y \in N_1 : \ Qx \in N_2(R(y,x))\},$$

For example, if Q is $\exists$, N_1 the set of women, N_2 the set of dogs, and R feeding, then this means there exists exactly 1 dog-feeding woman. If Q is *at least*

two, it says that there exists exactly 1 woman who feeds at least two dogs. It will be convenient to formalise quantifiers as relations between sets. Using $\forall$ as an example, instead of writing $\forall x \in A(x \in B)$, we write $\forall(A, B)$; more generally, we write $Q(A, B)$, where Q is any quantifier and A and B are sets of domain objects. Thus, the formula above is re-written as

$$\{y \in N_1 : Q(N_2, \{z : R(y, z)\})\},$$

where N_2 plays the role of A and $\{z : R(y, z)\}$ the role of B. Instantiating this as before, with $N_1 = Woman$, $Q = \exists$, $N_2 = Dog$, and $R = Feed$, we obtain $\{y \in Woman : \exists(Dog, \{z : Feed(y, z)\})\}$, or "women who feed a dog".

To show which quantifiers are expressible in our logic, let us think of quantifiers as quantitative constraints on the sizes of $A \cap B$, $A - B$, and $B - A$, as is common in the theory of Generalized Quantifiers. The findings are summarized in Table 10.1. $\mathcal{SROIQ}$ can express Logical Forms of the types in the table, plus disjunctions and conjunctions of such Logical Forms.

type	QAB	DL
1	$\geq n\ (Dog, \{z \mid Feed(y, z)\})$	$[y: \geq nFeed.Dog]$
2	$\geq n\ (Dog, \neg\{z \mid Feed(y, z)\})$	$[y: \geq n\neg Feed.Dog]$
3	$\geq n\ (\neg Dog, \{z \mid Feed(y, z)\})$	$[y: \geq nFeed.\neg Dog]$
4	$\geq n\ (\neg Dog, \neg\{z \mid Feed(y, z)\})$	$[y: \geq n\neg Feed.\neg Dog]$
5	$\leq n\ (Dog, \{z \mid F(y, z)\})$	$[y: \leq nFeed.Dog]$
6	$\leq n\ (Dog, \neg\{z \mid Feed(y, z)\})$	$[y: \leq n\neg Feed.Dog]$
7	$\leq n\ (\neg Dog, \{z \mid Feed(y, z)\})$	$[y: \leq nFeed.\neg Dog]$
8	$\leq n\ (\neg Dog, \neg\{z \mid Feed(y, z)\})$	$[y: \leq n\neg Feed.\neg Dog]$

Table 10.1

Expressing 8 types of Quantified REs *in Description Logic. Both columns of the table use "¬" to express set complementation as well as logical negation.*

Having all the quantifiers of $\mathcal{SROIQ}^+$ ($\mathcal{SROIQ}$ with role negation, see section 10.2) enables us to refer in a wide class of situations. When $n = 1$, type 1 in the table is equivalent to $\exists$. When $n = 0$, type 7 is equivalent to $\forall Feed.Dog$, that is, the quantifier *only*. When $n = 0$, type 6 becomes $\forall\neg Feed.\neg Dog$, that is, the quantifier *all*. In types 2, 4, 6, and 8, negation of a relation is used. This is not directly supported in $\mathcal{SROIQ}$, but, as indicated in section 10.2, given a relation $Feed$, its negation $\neg Feed$ can be used.

By logically conjoining the options offered by Table 10.1, these constructs allow the expression of a description such as "women who feed at least 1 and

at most 7 dogs", by conjoining type 1 (with $n = 1$) with type 5 (with $n = 7$). Exact numbers can be expressed as well, as in $\geq 1Feed.Cat \sqcap \leq 1Feed.Cat$ ("feed exactly one cat"), or intervals, as in $\geq 10Sell.Car \sqcap \leq 20Sell.Car$ ("sell between 10 and 20 cars"). In addition to Table 10.1, $\mathcal{SROIQ}$ can even represent relations to self, such as "the dog who loves itself" by $Dog \sqcap \exists Love.Self$, which was not expressible in the approach discussed in [Gardent and Striegnitz, 2007]. The usefulness of all these extensions is discussed in section 10.7.

Comparing the above with classes of quantifiers studied in the theory of Generalized Quantifiers, it is clear that $\mathcal{SROIQ}$ is highly expressive (cf., [van Benthem, 1986], chapter 2). Yet some quantifiers routinely expressed in English are not expressible in $\mathcal{SROIQ}$; examples include *most*, *infinitely many*, and *many*. These and other limitations of the above will be discussed in section 11.2. But first, let us see how our extended set of quantifiers enables us to generate a wider class of REs, making it possible to identify referents uniquely in situations where this was not possible before.

10.5 Using $\mathcal{SROIQ}^+$ to Generate Complex REs

Computational models are the theme of this book. But this term covers a multitude of sins. At one extreme lie models that have been implemented in computer programs and whose properties have been tested both mathematically (e.g., in terms of computational complexity) and empirically (e.g., in terms of their ability to mimic human reference production). At the other extreme lie models that are incomplete sketches, whose details have yet to be worked out, and where much about the algorithm is still unknown.

If we use the word *maturity* to talk about these differences, then the classic REG algorithms that form the theme of Part II of the book are positioned close to the more mature extreme of this range (despite remaining questions over, for example, the treatment of ties in the Full Brevity and Greedy algorithms; see our discussion of algorithm 3). Most of the algorithms discussed in the more recent chapters 8 and 9 lie somewhere in the middle of the range. The same is true of PRO, the probabilistic approach to REG defended in section 6.3: its core, which we dubbed the PRO *model* (Figure 6.4), has been evaluated extensively, but to make the model more widely applicable it was then extended to the PRO *algorithm* (algorithm 9), and this algorithm as a whole has not been implemented and evaluated yet.

The algorithm we are about to put forward lies at the less mature extreme of this range, where the treatment of proper names in chapter 2.7 lies as well. The increased logical expressivity that $\mathcal{SROIQ}$ offers, and that the REs of the previous section require, creates a computational "embarrassment of riches", not least in terms of keeping the algorithm computationally tractable (cf., section 4.8). The problem, in a nutshell, is the cardinal problem of Natural Language Generation: any given referent can be referred to in numerous, and highly varied, ways. Which of these are most likely to be uttered by a human speaker, and which are most useful for hearers, we do not know.

Our strategy, in the algorithm below, which we call GROWL after the language OWL2 to which $\mathcal{SROIQ}$ is closely related, is similar to that of Areces et al.: we generate REs for many subsets of the domain simultaneously, by computing satellite sets; like its predecessor, GROWL applies a generate-and-test strategy that composes increasingly complicated Logical Forms, without aiming to individuate one particular target referent. This time, however, the highly expressive Description Logic $\mathcal{SROIQ}$ is used and, unlike its predecessor, GROWL uses DL reasoning to find out what set is denoted by a Logical Form, taking TBox axioms into account. Moreover, as explained in the previous section, GROWL uses a large variety of quantifiers, instead of only $\exists$.

GROWL takes a Knowledge Base Σ as its input and outputs a queue LF of Logical Forms, from the syntactically simplest ones to increasingly complex ones. Different formulas can mean the same, for instance $\neg\forall R.A$ is equivalent with $\exists R.\neg A$. To reduce the set of formulas that our algorithm needs to consider, we look at Logical Forms in their *negation normal form* (NNF). A NNF has $\neg$ in front of only atomic concepts (including $\top$ and $\bot$), or self-restrictions. The NNF of $\neg C$ ($\neg R$) is denoted ~C (~R). The algorithm uses these sets:

The set RN of atomic relations. For any atomic relation R, RN contains R, ~R, R^-, and ~R^-.

The set CN of atomic concepts. In addition to concepts such as *Man*, *Woman*, etc., CN contains $\top$ and $\bot$ and all expressions of the form $\exists R.Self$ where $R \in RN$. For any atomic concept A, CN also contains the NNF (~A).

A finite set N of integers $1, \ldots, n$, where n does not exceed the number of individuals in Σ. N bounds the number n used in numerical quantifiers.

The construct set Con containing all the constructs whose use in Concept expressions $\mathcal{SROIQ}$ permits: $\{\neg, \sqcap, \sqcup, \exists, \forall, \leq, \geq, =\}$.

Before we sketch the GROWL algorithm, note that square brackets indicate strings of symbols. For instance, in Step 8, where d and d' denote elements of

LF, and s denotes a set-theory connective, the expression "$[d\ s\ d']$" denotes the string that intersects, or forms the union of, two elements of LF.

Steps 2-3 add suitable elements of CN to the set LF, which starts off empty. The details of Add can be specified in different ways; [Ren et al., 2010] uses a simple heuristic, whereby new Logical Forms are only added if they have smaller extensions than all elements of the existing LF, so if two descriptions are equivalent, then at most one of them ends up in LF. Crucially, these procedures take implicit knowledge (i.e., TBox axioms) into account.

Algorithm 19 GROWL: Computation of $\mathcal{SROIQ}^+$-satellite sets

Input: The Knowledge Base Σ and the sets CN, RN, N, and Con. An interpretation function $\mathcal{I}$ that interprets CN and RN.
Output: A queue LF of formulas φ such that $\{\varphi^{\mathcal{I}} : \varphi \in \text{RE}\}$ is the set of things that are $\mathcal{SROIQ}^+$-satellite sets (of any object in $\Delta^{\mathcal{I}}$) given $\mathcal{I}$.

```
1:  LF := ∅
2:  for each e ∈ CN do
3:      LF := Add(LF, [e])
4:  for each d = fetch(LF) do
5:      for each s ∈ Con do
6:          if s = ⊓ or s = ⊔ then
7:              for each d' ∈ LF do
8:                  LF := Add(LF, [d s d'])
9:          if s = ∃ or s = ∀ then
10:             for each r ∈ RN do
11:                 LF := Add(LF, [s r.d])
12:         if s = ≤ or s = ≥ or s is = then
13:             for each r ∈ RN and each k ∈ N do
14:                 LF := Add(LF, [s k r.d])
15: return LF
```

During Steps 4-14, elements of LF are processed recursively, one by one. $fetch(LF)$ retrieves the first element of LF, then the next, and so on. New elements are added at the end of LF and are never removed. For each element d of LF, Steps 6-14 generate a more complex Logical Form and add it at the end of LF. Let us call this the *expansion* of LF. Steps 7 and 8 conjoin or disjoin d with each element of LF. Steps 10 and 11 extend d with $\exists$ and $\forall$, then Steps 13-14 extend d with numerical quantifiers. During the expansion of LF, GROWL does not consider $s = \neg$ because, in NNF form, negation appears only in front of atomic concepts.

At any stage of the execution of GROWL, LF, RN, N, and S are finite, hence Steps 5 to 14 terminate for a particular $d \in LF$. Assuming a finite domain, the algorithm as a whole terminates as well, because of the way in which *Add* is constructed (i.e., at some point there are no new subsets of the domain left to be referred to). Consequently, GROWL can be used in the same way as the algorithm of Areces and colleagues (section 10.3), generating one description for each satellite set. If the aim is to find a description for a specific target referent only, then construction of LF could be stopped once a description of the target referent has been found.

The pseudo-code of algorithm 19 embodies an implicit "theory" that the connectives $\sqcap$ and $\sqcup$ add less to the complexity of a description than the quantifiers $\exists$ and $\forall$, which add less than the numerical quantifiers. Some decisions are left unspecified, however, leaving it open whether $\forall$ or $\exists$ is added first, for instance. Once these decisions are made, GROWL generates at most one description for a given referent. If we apply the algorithm to the Knowledge Base in Figure 10.2, the following solutions can be generated:

1. $\{w1\} = Woman \sqcap \exists\neg feed.Cat$, the woman who does not feed all cats
2. $\{w2\} = Woman \sqcap \leq 0\neg feed.Cat$, the woman who feeds all cats
3. $\{d1\} = Dog \sqcap \leq 0\neg feed^{-}.Woman$, the dog that is fed by all women
4. $\{d2\} = Dog \sqcap \exists\neg feed^{-}.Woman$, the dog that is not fed by all women
5. $\{c1\} = Cat \sqcap \leq 0\neg feed^{-}.Woman$, the cat that is fed by all women
6. $\{c2\} = Cat \sqcap \exists\neg feed^{-}.Woman$, the cat that is not fed by all women

In other words, where previous algorithms were unable to refer to even one object in this domain, GROWL is able to refer to all six.

Like the other algorithms discussed in this chapter, GROWL focusses on finding REs. Ultimately, the algorithm should be fine-tuned based on empirical research. For example, a TUNA-style evaluation (chapter 5) could be applied to domains in which relations (like feed, love, etc.) play a role, to see how people use quantifiers when they refer. Alternatively, a hearer-oriented type of evaluation (perhaps in the style of the GIVE challenge, section 5.8) could be performed to find out what types of REs permit hearers to individuate the referent most speedily and reliably.

10.6 Rethinking REG: Using Shared Knowledge That Is Not Atomic

We have discussed the classic REG task (section 4.3) and various extensions of this task. From where we stand now, it is possible to see further than before, towards the generalizations that will be discussed in the next Part of the book. One extension, however, is so closely related to the subject matter of the present chapter that we discuss it here.

Representations of shared knowledge in REG have long been limited to atomic facts and their implicit negations; more complex types of shared knowledge have rarely been considered. Non-atomic formulas often permit a more succinct and insightful representation of information. But while succinctness and efficiency are laudable, are they crucial? It might be thought that after the reasoner has done its work, a Knowledge Base in the familiar style results, as in section 10.4 when the model of Figure 10.2 resulted from reasoning with the one in Figure 10.1. Consequently, we might expect any reasoning to have been done before the start of REG, using a *pre-compilation* strategy that computes all the instances of each atomic concept, and all the pairs of domain objects in each atomic relation, once and for all. (The result is known as the *materialized* ABox.) However, complete pre-compilation is not always feasible.

The main reason why pre-compilation does not always work is the kind of incomplete knowledge flowing from disjunctions and existential quantifiers. Consider our earlier example of a painting that is known to be Flemish or Dutch, say $[a : Flemish \sqcup Dutch]$. We also know that a certain other painting is Spanish, say $[b : Spanish]$. This is all we know; hence the Knowledge Base has models in which a is Flemish and models in which a is Dutch. We cannot simply let the reasoner compute a set of atomic propositions, because neither $[a : Flemish]$ nor $[a : Dutch]$ is justified. It is possible to refer to a as "The painting that is Flemish or Dutch" (or "The painting that is not Spanish"), but the pre-compilation strategy outlined above will end up empty handed. GROWL, by contrast, is able to generate a disjunctive RE that does the job.

It may be worth reflecting on the nature of the knowledge that we want to model. One might argue that a Knowledge Base should be able to "decide" each and every atomic proposition in the language. According to this perspective, a painting may be Flemish, it may be Dutch, but it cannot be Flemish-or-Dutch. This perspective on knowledge is embodied in the classic REG algorithms of Part I, and in the algorithm of Areces and colleagues. But if we want to model human language production, then the objective perspective is

hopelessly limited. In many situations, a person's knowledge is incomplete: the information state of someone who looks at the painting, uncertain as to whether it is Dutch or Flemish, is characterized best by the disjunctive statement $[a : Flemish \sqcup Dutch]$. This is one of the reasons why a genuine Knowledge Base is a better model of human knowledge than a simple database.

An example due to Yuan Ren shows that pre-compilation can even result in erroneous REs. Consider the following ontology, which hinges on incomplete knowledge in the form of an existential quantifier. Suppose the only ABox axioms in the ontology are $[a : A]$, $[b : A]$, and $[(a, b) : R]$, and the only TBox axiom is $A \sqsubseteq \exists R.A$ (every element of A stands in the relation R to something in A). Materialization has no effect: the materialized ABox is the same as the original one. Using a REG algorithm directly on the ABox would result in an RE that singles out a as the only domain object that stands in the relation R to something in A (suggesting that it is the only instance of $\exists R.A$). However, by consulting the TBox axiom, GROWL will find out that b is an instance of this concept as well, so the previous RE is incorrect. Likewise, a reliance on pre-compilation would cause b to be erroneously referred to as the only element of $\neg\exists R.A$, whereas in fact it is in $\exists R.A$. Materialization simply does not do justice to an ontology expressed in $\mathcal{SROIQ}$.

Other cases of reasoning intrigue for philosophical reasons (cf., section 2.6). Suppose we wanted to generate an RE like "the last quarterly report of 2009" from a Knowledge Base that makes optimal use of generic rules, saying that every year has a unique first, second, third, and fourth quarter, and that every quarter has a unique report associated with it. The reasoner will deduce that there exists a unique report on the fourth quarter of 1929, allowing the generator to refer to it producing REs such as "the last quarterly report of 1929". Suppose now we are interested how a certain company weathered the economic crisis of the late 1920s. October 1929 was the crucial month for this company, so the last quarterly report should contain the information we are looking for. We may want to say,

(U) The report on the last quarter of 1929 must be crucial.

But, we do not have any direct acquaintance with this report, arguably making "The report on the last quarter of 1929" an *attributive* description in Donnellan's sense (cf., section 2.6). To see this, we can apply a standard test to distinguish between attributive and referential descriptions: suppose you had uttered U when you first saw a volume the librarian had fetched for you. When

Figure 10.4
Attributive description: "What a mess! The murderer *must have been insane."*

you uttered U, you believed this volume to contain the report of the last quarter of 1929, but now you open it, and you discover that it's actually a volume of poetry. This discovery tells you that the referent of "the report of the last quarter of 1929" is not what you thought it was. Yet it will not make you to want to revise U, because the utterance itself remains accurate. This thought experiment suggests that your utterance U was not about a particular thing; rather, it expressed a general rule, facilitated by an attributive description that means, approximately, "whatever volume reports on the last quarter of 1929". Donnellan's original examples of attributive descriptions can be modelled in analogous fashion. Consider his

The murderer of Smith is insane,

assuming the scenario displayed in Figure 10.4: Smith's mortal remains, in the midst of a scene of utter devastation, are tied in a knot. To represent the relevant information, a Knowledge Base could contain concepts such as Corpse, Murderer, Being tied in a knot, and Being insane. Its ABox would state the observable facts, whereas the TBox might state that anyone tied into a knot

has been murdered by exactly one murderer, and that any murderer who ties her victim into a knot is insane. Such an ontology would permit the attributive description, "the murderer of Smith".

This analysis is not without its rough edges, for example because many rules – such as the one about the existence of quarterly reports – permit exceptions; after all, 1929 may have been such a troublesome year that one quarter went unreported. Nonetheless, automated reasoning has the capacity to expand REG algorithms into areas that have long been studied by theoreticians, but that approaches based on a database of atomic statements had never been able to reach. The new approach does not make the problems with attributive descriptions disappear overnight, but they do open a new, constructive perspective on the matter that might bear fruit.

The approach to REG outlined in this section offers a stark counterpoint to Dale and Reiter's move (section 4.3) away from the complexities addressed by the California School of REG (section 4.2): if REG is approached as outlined in the present section, then REG is becoming even more complex than it was in the 1980s. Once again, there would be a need for extensive modelling of knowledge. Of course, the modelling of knowledge has advanced very considerably since the 1980s, not least as a result of work on Description Logic, with its ability to perform a lot of reasoning tasks very rapidly, so researchers wanting to work in this direction will have a much better starting point than those in the California School. Maybe what was too difficult in the 1980s has become possible now.

Lessons may be learned from projects such as CYC, which formalize common-sense knowledge [Matuszek et al., 2006]. This type of detailed logical modelling, sometimes known as "real AI", does not sit easily with current trends in Computational Linguistics (discussed in the introduction to chapter 5), because real AI is labour intensive and domain dependent. Perhaps the way forward is to acquire models of common-sense knowledge automatically, an enterprise into which considerable effort is being invested at the moment, often based on Machine Learning from text (e.g., [Ryu et al., 2010]). This, to me, seems an interesting way forward. Needless to say, the simpler perspective on REG on which this book focusses will suffice for producing REs in many practical applications of Natural Language Generation. But if the aim is to achieve the fullest possible understanding of speakers' ability to refer – the construction of computational models of referring, in other words – then common-sense knowledge should not be overlooked.

10.7 Why Study the Generation of Logically Complex REs?

The previous chapters have shown how the expressive power of REG has grown. But how useful are these extensions? Might they be like Rube Goldberg machines, whose pay-offs are negligible in comparison with the complexity of their mechanisms?

With a few notable exceptions, computational linguists have recoiled from their early fascination with logic, as we have seen (e.g., section 4.2, the introduction to chapter 5, and section 10.6). In this intellectual climate, the relevance of modern Knowledge Representation will not be taken for granted. There are good reasons for this skeptical attitude, not least because experimental psychology has taught us that our species' grasp of logical inference is tenuous (e.g., [Johnson-Laird, 2006], [Kahneman, 2012]); it might therefore be thought that there cannot be a place for the complexities of Formal Logic in an account of human speaking. Yet I will argue that it is important to study logical extensions of REG and the questions that they raise. This time around, however, Formal Logic should go hand in hand with empirical *finesse*, based on corpora and experiments with human language users; empirical sophistication, as well as speed and expressivity, were lacking in the 1980s (section 4.2).

1. **Sometimes the referent could not be identified before.** Sometimes the new algorithms allow us to individuate a referent where this was previously impossible. We have used this argument time and again: referents that were not referable for the classic REG algorithms have become referable because of later extensions in the expressive power of these algorithms.

2. **Sometimes the proposed extensions generate less complex REs.** Some of the new REs are relatively simple and occur frequently. This goes without saying for proper names (section 7) and for quantifiers like "all" ("The woman who programmed all these functions"), "two" ("The dog with two owners"), and "only" ("The dog that only chases cats").

People may have difficulties with some of the constructs employed by the new algorithms: our cognitive difficulties with negation are well attested, for instance. Still, an RE that contains a negation may be *less* complex than any other RE that singles out the referent. Consider a car park full of vehicles. Given the choice between "The cars that are not Hondas need to be removed" and "The Fords, Toyotas, Audis, (...), need to be removed", the former may be shorter and more preferable. My claim is not that negation should be used at

every opportunity, but that in some situations, REs that use negation are preferable to others. This argument applies to quantified REs as well. Expressions like "The man who adores *all* cars" and "The woman with *two* suitors", may be complex, yet they will sometimes be simpler than the simplest non-quantified RE that singles out the referent.

3. **Simplicity isn't everything.** So far, we have tacitly assumed that a complex Logical Form is worse than a simple one, suggesting that a generator should always prefer the simplest one. But simplicity is not everything. Consider

– "The man who loves all cars" (when a car exhibition is discussed)
– "John's car" (when John is salient or after discussing someone else's car)
– "The cars that are not Hondas" (in front of a Honda factory).

In all these situations, the RE contains a source of complexity (the quantifier "all", a relation, or a negation) for which it should be rewarded rather than penalized, because it makes the RE more contextually relevant than a simpler RE would have been (cf., section 4.2).

4. **A complex content does not always require a complex form.** We have been assuming that Content Determination precedes decisions about the syntactic and lexical form of the RE. The English expression coming out at the end of the generation pipeline may not always mirror the logical structure of the Logical Form directly: a complex formula can sometimes be expressed through simple words. For example,

$\geq 10^6 Own.Dollars$ ("The person who owns at least a million dollars"): "The millionaire"

Furthermore, the information in the RE might be dispersed over several utterances (cf., section 3.6). Suppose the generator produces $Woman \sqcap \exists Feed.(Dog \sqcap \forall Chase.Cat)$. Parts of this RE may end up scattered over various dialogue turns, the first of which might describe someone as (1) "This woman", the second might add that (2) "she feeds a dog", and the third says that (3) "the dog chases only cats". By breaking complex information down into bite-size chunks, the RE as a whole may well become easier to understand. Further research is needed to spell out the logical constraints on the dispersal of referential information, which have to do with the monotonicity properties of the different types of quantifiers. To see the problem, consider the information in a slightly different RE, $Woman \sqcap \leq 0 \neg Feed.(Dog \sqcap \forall Chase.Cat)$.

A direct translation might say "the woman such that there are no dogs that chase only cats whom she does not feed"; after some logical manipulation, this becomes "the woman who feeds all dogs that chase only cats". The information in this RE cannot, however, be dispersed over three separate utterances, as in (1) "This woman", (2) "she feeds all dogs", (3) "the dogs chase only cats", because the second utterance may be false.

5. **Characterizing linguistic competence.** Finally, exploring the space of possible REs is of theoretical interest. Let us assume, against our better judgment, that empirical research will show beyond reasonable doubt that most of the newest batch of REs are useless: no human speaker would ever generate them, and no human hearer would ever benefit from them. I would contend that our understanding of human language and language use would be considerably advanced by this insight, because it would show us which logically possible referential strategies people actually use.

Two famous analogies come to mind. First, consider center-embedding, for example of relative clauses. As is well known, this syntactic phenomenon can give rise to arbitrarily deep nestings, which are easily covered by recursive grammar rules yet difficult to understand (e.g., "a man that a woman that a child knows loves") [Karlsson, 2007]. There exists a finite upper bound on the depth of embedding of any NP ever encountered, therefore a grammar without recursive rules can cover all NPs. Does this mean there is no need for recursive rules? Probably not. A standard response, ever since Chomsky's [Chomsky, 1965], is that recursive rules are needed for modelling the human linguistic *competence* (i.e., what people can say in principle); human linguistic *performance* may be best understood by additional constraints, reflecting limitations on human memory and calculation. The semantically complex expressions of the present chapter may be viewed in the same light.

The second analogy transports us back to our discussion of quantifiers. The theory of Generalized Quantifiers does not stop at a characterization of the class of all quantifiers. The theory also asks what quantifiers can be expressed through syntactically simple means (e.g., in one word, or within one NP [Peters and Westerstahl, 2006]). It is this theory that enabled us to extend REG algorithms as proposed in section 10.4. Once we know what REs are possible, we can ask which of these are actually used. In this way, the study of reference production opens up a new chapter in the study of Generalized Quantifiers, focussing specifically on the use of quantifiers in referring expressions.

10.8 Summary of the Chapter

This chapter has argued that REG can benefit greatly from the expressive power and reasoning support that modern Knowledge Representation paradigms such as Conceptual Graphs and Description Logic can offer. Specifically, we have shown how Description Logic has started to extend the power of REG significantly, and we have outlined how further extensions are possible.

- Areces and colleagues have shown that the Description Logics $\mathcal{ALC}$ and $\mathcal{EL}$ can be employed to generate REs in an elegant and a principled way. This is true for "classic" REs in the style of Dale and Reiter, for relational REs, and for REs that make use of negated properties. [Section 10.3]
- Substantial advantages are obtained when more highly expressive logics such as $\mathcal{SROIQ}$ are employed, permitting the identification of entities that no earlier REG algorithm was able to identify, containing properties such as "feeds two cats" and "is not loved by all dogs". It appears that $\mathcal{SROIQ}^{+}$, a minor extension of $\mathcal{SROIQ}$, permits REG to use those quantifiers that are most important for natural language. [Sections 10.4 and 10.5]
- If REs are generated from logically structured information, this opens up a range of new algorithmic possibilities. For example, it enables the exploitation of incomplete knowledge, as when we say "The painting that is Dutch or Flemish". Also, it would permit the generation of *attributive* descriptions in the style of Donnellan, as when we speculate about "the report on the last quarter of 1929". Thus it will enable our algorithms to talk about entities that are not explicitly mentioned in the Knowledge Base but whose (unique) existence can be deduced. [Section 10.6]
- It is practically useful to study negation, disjunction, and quantifiers because without these operations, unique reference is often impossible. It is also theoretically important to study these operations because this gives us a starting point for investigating why some REs come to human speakers more naturally than others. [Section 10.7]

Having used the bulk of Part III of this book to discuss ways in which REG algorithms can be made more expressive and powerful, we shall use the final chapter of this Part to reflect on their limits. After all the extensions discussed here, are REG algorithms able to identify any referent that can be identified at all, or might further extensions be called for?

11 The Question of Referability

In the last few chapters, we have seen algorithms that produce a distinguishing description in situations where algorithms designed in the 1980s and 1990s fall short. The limitations of earlier algorithms arose from an inability to represent negation, disjunction, and certain quantifiers. So where are we now? Are we able to generate a distinguishing description whenever one exists? We shall see that the answer to this question is: not quite.

In this short chapter, as in the previous one, we focus on *logical* completeness alone: syntactic and lexical completeness are well beyond our grasp (as we saw in section 1.1). We shall focus on reference to a single individual, but generalizations to sets will be straightforward. We shall start by asking what it might mean for a target referent to be "unreferable", and this will lead to a Referability Theorem, which offers a semantic test of whether two domain objects can be told apart by any predicate-logical formula (section 11.1). Next, we investigate some implications of this theorem, concluding that even the GROWL algorithm, which we sketched in the previous chapter, is unable to refer in all situations in which reference is possible (section 11.2). We conclude the chapter, and the Third Part of the book, by asking how REG algorithms might go beyond GROWL (section 11.3).

11.1 Revisiting the Logical Completeness of REG

In section 4.7 I wrote: "Let's call a REG algorithm *successful* with respect to a given situation, characterized by a Knowledge Base and a given target r, if the algorithm produces a distinguishing description in that situation. We will call an algorithm (logically) *complete* if it is successful in every situation in which a distinguishing description exists." Henceforth, we shall say that if a distinguishing description of r exists, then r is *referable*.

It would be unfair to complain that the classic REG algorithms (or the GROWL algorithm, for that matter) is unable to exploit epistemic modalities, or probability for example, because these types of information are not represented in the Knowledge Base that forms the starting point for these algorithms. In what follows, we shall show that there does exist an absolute sense in which an individual may be referable, and in which a REG algorithm may be logically complete, provided some assumptions are made concerning the information expressed in the Knowledge Base. To the best of my knowledge, these assumptions are met by any Knowledge Base to which REG algorithms have so far been applied.

In chapter 10, where we used Description Logic, the notion of an *interpretation* (also called a model) was defined as a pair $\langle \Delta^{\mathcal{I}}, .^{\mathcal{I}} \rangle$, where $\Delta^{\mathcal{I}}$ is a non-empty set and $.^{\mathcal{I}}$ is a function that maps each atomic concept to a subset of $\Delta^{\mathcal{I}}$, each atomic relation (i.e., role) to a subset of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, and each individual to an element of $\Delta^{\mathcal{I}}$. Implicit in this definition is the idea that individuals belong to concepts (i.e., properties) and are connected to each other by 2-place relations. Let us assume our present models to be exactly like that, disregarding functions, and relations of higher arity. In connection with REG, it is reasonable, moreover, to assume that models have only finitely many entities. The models that result from these assumptions can be graphically depicted as in the diagrams of the previous chapter. Under these assumptions, what would it mean for a REG algorithm to be logically complete?

Indistinguishable entities. Is it ever possible to prove that one domain element cannot be distinguished from another one? More generally, is there such a thing as "the" satellite set of an entity, regardless of the logical language under consideration? Because negation can be expressed in some logics (hence, if a formula φ distinguishes a from b, then $\neg\varphi$ distinguishes b from a), we can make the question symmetrical, asking under what conditions two entities cannot be distinguished *from each other*.

In real life, it can be extremely challenging, or even impossible, to find an effective way to individuate a referent. (See Figure 11.1, where the speaker is at a loss to produce a distinguishing description of her backpack, because it is very similar to other backpacks.) In strictly logical terms as well, some models contain elements that are indistinguishable from each other. Consider a model that contains two men and two women, happily arranged as follows:

$[m_1 : Man], [w_1 : Woman], [(m_1, w_1) : Love]$
$[m_2 : Man], [w_2 : Woman], [(m_2, w_2) : Love]$

Logically there is nothing to separate the two men, or the two women for that matter. The model has two "halves" that are isomorphic to each other (i.e., there exists a 1-1 mapping between the two halves that respects all properties and relations). The same is true in a less fortunate variant of the model that has only one woman, loved by both men:

$[m_1 : Man], [m_2 : Man], [w_1 : Woman]$
$[(m_1, w_1) : Love], [(m_2, w_1) : Love]$

Figure 11.1
Reference is problematic if the referent is indistinguishable from one or more of its distractors.

This example shows overlapping halves (because w_1 belongs to each half) which are nonetheless isomorphic, capturing the intuition that there is nothing to separate m_1 and m_2. In this example and the previous one, m_1 and m_2 cannot be distinguished from each other. These observations suggest that two objects can only be told apart if they take up different parts in the "topology" of the model.

In formalizing this idea, it will sometimes be convenient to disregard the direction of the arrows in our graphs, viewing them as if they were undirected arcs. Given an entity r in M, we now define the model *generated* by r, abbreviated $M(r)$, as the (directed) part of M that is reachable from r. More precisely, and continuing to limit ourselves to relations whose arity does not exceed 2, let $M(r)$ be the result of restricting the model M, with its set of individuals $\Delta^{\mathcal{I}}$, to the set $Reach(\Delta^{\mathcal{I}}, r)$ (the subset of $\Delta^{\mathcal{I}}$ reachable from r) consisting of all those objects in $\Delta^{\mathcal{I}}$ to which one can travel from r following the arcs regardless of their direction, and including the starting point r itself: (Note the recursion in Steps 2 and 3.)

1. $r \in Reach(\Delta^{\mathcal{I}}, r)$.
2. For all objects x and y and for every relation R,
if $x \in Reach(\Delta^{\mathcal{I}}, r)$ and $(x, y) \in R^I$, then $y \in Reach(\Delta^{\mathcal{I}}, r)$.
3. For all objects x and y and for every relation R,
if $x \in Reach(\Delta^{\mathcal{I}}, r)$ and $(y, x) \in R^I$, then $y \in Reach(\Delta^{\mathcal{I}}, r)$.
4. Nothing else is an element of $Reach(\Delta^{\mathcal{I}}, r)$.

We prove that, under the stated assumptions, an object is referable if and only if it can be referred to by means of a formula of First-Order Predicate Logic With Identity (FOPL). The key is the insight that two elements, a and b, are indistinguishable using FOPL if and only if the models $M(a)$ and $M(b)$ generated by a and b are isomorphic, with a and b taking up analogous positions in their respective submodels. We write $[\![\varphi(x)^M]\!]$ to denote the set of objects in M (i.e., in $\Delta^{\mathcal{I}}$) that satisfy the FOPL formula $\varphi(x)$. The proof will use the fact that, given a submodel M' generated by one of the objects in M, a FOPL formula describing M' completely up to isomorphism can be "read off" M', as follows:

Reading off a Logical Form for the referent r from a generated model M':

1. Logically conjoin all atomic propositions p for which $M' \models p$.
2. Add, as additional conjuncts, inequalities $a \neq b$ for all constants a, b occurring in the conjunction resulting from (1).
3. Close off all 2-place relations. Thus, for any a, if a_1,..,a_n are the only constants a_i such that $M' \models aRa_i$, then add, as an additional conjunct, the formula $\forall y(aRy \rightarrow (y = a_1 \vee ... \vee y = a_n))$; similarly, if a_1,..,a_n are the only constants a_i such that $M' \models a_iRa$, then add, as an additional conjunct, the formula $\forall y(yRa \rightarrow (y = a_1 \vee ... \vee y = a_n))$.
4. Replace all occurrences of constants by free variables (using the same variable for occurrences of the same constant, and different variables for occurrences of different constants), using x to replace r. Occurrences of the same constant are replaced by the same variable, occurrences of different constants by different variables.
5. Quantify existentially over all the free variables in the conjunction, except for the variable x, which remains free.

The third step is reminiscent of *circumscription* [McCarthy, 1980]: this step makes it explicit that relations hold only where M' says they hold. To see why this step is required, suppose M is as follows, and our intended referent is m_1:

$[m_1 : Man]$, $[w_1 : Woman]$, $[(m_1, w_1) : Love]$
$[m_2 : Man]$, $[w_2 : Woman]$, $[w_3 : Woman]$,
$[(m_2, w_2) : Love]$, $[(m_2, w_3) : Love]$.

In this model, m_1 cannot be identified without taking into account that m_1 does *not* love w_2 and w_3. Let's go through the five steps above, omitting 1-place predicates for simplicity. Starting from the generated model $M(m_1)$, Step **1** yields a short formula, $L(m_1, w_1)$. Step **2** turns this into the conjunction $L(m_1, w_1) \wedge m_1 \neq w_1$. Step **3** adds two universally quantified conjuncts, resulting in $L(m_1, w_1) \wedge m_1 \neq w_1 \wedge \forall y(L(x, y) \rightarrow y = w_1) \wedge \forall y(L(y, w_1) \rightarrow y = m_1)$. Step **4** replaces m_1 by the free variable x and replaces w_1 by some other variable, say x_1. Step **5** add an existential quantifier as a prefix, yielding $\exists x_1(L(x, x_1) \wedge x \neq x_1 \wedge \forall y(y \neq x_1 \rightarrow \neg L(x, y)) \wedge \forall y(L(y, w_1) \rightarrow y = x))$. This formula says that x loves one and only one person, that this person is not x, and that this person is loved by x only. This formula is true of m_1 and of no other element of M.

We are now ready state and prove a key Lemma:

Lemma 2 Consider a finite model M that contains elements a and b. Now a and b are FOPL-*indistinguishable* if and only if there exists a bijection f from $M(a)$ to $M(b)$, such that f respects all properties and relations and such that $f(a) = b$. We sketch the proofs of the two directions:
$\Longleftarrow$ Let f be a bijection from $M(a)$ to $M(b)$ as specified. Then one can prove using formula induction that $a \in [\![\varphi(x)^M]\!]$ iff $b \in [\![\varphi(x)^M]\!]$, for any FOPL formula $\varphi(x)$ (i.e., a and b are FOPL-*indistinguishable*). Note that this equivalence holds across all of M: we do not merely prove that $a \in [\![\varphi(x)^{M(a)}]\!]$ iff $b \in [\![\varphi(x)^{M(b)}]\!]$.
$\Longrightarrow$ Suppose there does not exist f as specified. Then there exist a FOPL formula $\varphi(x)$ such that $\varphi(a)$ is true and $\varphi(b)$ is false (hence a and b are not FOPL-*indistinguishable*), and conversely. A suitable formula $\varphi(x)$ can be *read off* $M(a)$ in the manner specified above.

We have seen what it takes for an object to be distinguishable from another. Given a model, an intended referent r will be called FOPL-referable if there exists a FOPL formula that is true of r and false of all distractors, with respect to the model. The Referability Theorem follows almost immediately, because to be referable is to be distinguishable from every distractor.

Theorem 5 Referability Theorem: Consider a finite model M that contains an intended referent r and n distractor objects, $d_1, .., d_n$. Now r is *not* FOPL-*referable* if and only if, for some distractor d_i, there exist a bijection f_i from $M(r)$ to $M(d_i)$, such that f_i respects all properties and relations and such that $f_i(r) = d_i$. *Proof:* If there exists such a distractor d_i, then it follows from Lemma 2 that r is not FOPL-referable. If there does not exists such a distractor d_i, then the formula read off $M(r)$ shows r to be FOPL-distinguishable from each of $d_1, .., d_n$, and hence the same formula shows r to be FOPL-referable. □

Saying this informally, FOPL can characterize precisely those domain elements that take up a unique place in the model. Therefore, given our assumptions about the nature of a model, if FOPL cannot refer to a domain element, then this element cannot be referred to at all. Given the class of models specified, things that FOPL cannot identify cannot be identified at all.

The Referability Theorem might seem unsurprising, yet its details are worth taking in. For example, equality ($=$) is a crucial part of FOPL because without equality, one cannot distinguish between (on the one hand) an object that stands in the relation R to one thing and (on the other hand) an object that stands in the relation R to two different things.

The assumption that M is finite is crucial too, because for infinite models the second half ($\Rightarrow$) of the Lemma does not always hold. To see this, consider the counterexample of a large model M in which there are no properties, R is the only 2-place relation, and a and b take up similar positions in the graph of the model, with $\forall x(aRx \leftrightarrow x \in \{a_i : i \in \mathbb{R}\})$ (where $\mathbb{R}$ is the set of Real numbers, which is uncountable) and $\forall x(bRx \leftrightarrow x \in \{b_i : i \in \mathbb{N}\})$ (where $\mathbb{N}$ is the set of Natural numbers, which is infinite but countable), so b is R-related to only *countably* many objects, whereas a is R-related to *uncountably* many objects. Now the generated models $M(a)$ and $M(b)$ are clearly not isomorphic, yet no FOPL formula φ exists such that $\varphi(a) \wedge \neg\varphi(b)$ or $\varphi(b) \wedge \neg\varphi(a)$. This may serve as a reminder of the limitations of FOPL.

Note that the proof of the second half of the Lemma ($\Rightarrow$) is *constructive*. This suggests the possibility of new REG algorithms, which is an issue we shall turn to very briefly in section 11.3. The procedure of reading off a formula from a sub-model is reminiscent of the idea of a Satellite set (see sections 8.3 and 10.3): because the formula read off the sub-model generated by an entity r spells out all that is known about r, this formula does not only represent a logically exhaustive attempt at distinguishing this entity from one particular distractor; it distinguishes r from every other entity in the model from which it can be distinguished at all: if the formula happens to be true for any other entity in the model, then this other entity cannot be distinguished from r, hence it is a member of the Satellite set of r.

Our exposition in this section has been limited to properties and 2-place relations, but various generalizations can be proven straightforwardly. If models include relations of higher arity, then the concept $Reach(\Delta^{\mathcal{I}}, d)$ can be redefined to include all objects that can be reached through n-ary relations. (For instance, to cover 3-place relations, if $x \in Reach(\Delta^{\mathcal{I}}, d)$ and $(x, y, z) \in r^I$,

then $y \in Reach(\Delta^{\mathcal{I}}, d)$ and $z \in Reach(\Delta^{\mathcal{I}}, d)$). The definition of "reading off" should be reformulated analogously, after which the proof of the Referability Theorem proceeds in the same way as before. The theorem can also be generalized to include references to sets (cf., chapters 8 and 10 for algorithms that generate references to sets).

11.2 Limitations of $\mathcal{SROIQ}^+$ and the GROWL Algorithm

In recent years, REG algorithms have made great strides forward, not least in terms of their expressive power. Yet even the GROWL algorithm is not logically complete: there can be referable entities that GROWL is unable to identify.

One class of limitations is caused by the fact that $\mathcal{SROIQ}^+$ lacks full equality. For even though relations to self are expressible in $\mathcal{SROIQ}^+$, as in $Dog \sqcap \exists Love.Self$ ("dog that loves itself"), and even though the cardinality quantifiers imply an ability to tell entities apart, other uses of equality cannot be expressed, and this affects what one can refer to. Consider, for instance, the following model:

$[m_1 : Man]$, $[m_2 : Man]$, $[d_1 : Dog]$,
$[(m_1, d_1) : Feed]$, $[(m_2, d_1) : Feed]$, $[(d_1, m_1) : Love]$.

$M(m_1)$ is not isomorphic to $M(m_2)$, so both m_1 and m_2 are referable. For example, m_1 is the man who feeds dogs that love him. This is expressible in FOPL, but not in $\mathcal{SROIQ}^+$. In fact, there is no $\mathcal{SROIQ}^+$ concept that identifies m_1. In other words, m_1 is referable, yet the GROWL algorithm is unable to produce an RE that accomplishes this.

A second class of limitations is caused by the fact that $\mathcal{SROIQ}^+$ does not allow intersection between roles (i.e., relations). Consider a situation discussed by Gardent and Striegnitz, with a model M that contains two men and three children. One man adores and criticizes the same child; the other adores one child and criticizes the other [Gardent and Striegnitz, 2007]:

$[m_1 : Man]$, $[m_2 : Man]$, $[c_1 : Child]$, $[c_2 : Child]$, $[c_3 : Child]$,
$[(m_1, c_1) : Adore]$, $[(m_2, c_2) : Adore]$,
$[(m_1, c_1) : Criticize]$, $[(m_2, c_3) : Criticize]$

$M(m_1)$ is not isomorphic to $M(m_2)$, so m_1 and m_2 are distinguishable from each other in FOPL. A suitable formula φ true for m_1 but false for m_2 can be read off $M(m_1)$. Part of this formula says that $\exists x(Adore(m_1, x) \wedge Criticize(m_1, x))$. This information about m_1 would be false of m_2; hence, FOPL with equality can separate the two men. Can GROWL?

The quasi-formula $Man \sqcap \exists Adore \sqcap Criticize.Child$ ("the man who adores and criticizes one and the same child") comes to mind, but intersections of relations are not supported by OWL2, so this is not an option for GROWL. OWL2 can express $Man \sqcap (\exists Adore.Child) \sqcap (\exists Criticize.Child)$, but, in the model at hand, this denotes a set of two (i.e., $\{m_1, m_2\}$). OWL2, and hence GROWL, is unable to distinguish m_1 and m_2 from each other.[1]

These limitations of the GROWL algorithm stem from $\mathcal{SROIQ}$ rather than from the algorithm itself. $\mathcal{SROIQ}$ adds the limitation that it is only able to take cardinalities up to a certain size into account, which can prevent it from identifying entities in larger models. (The same holds *a forteriori* for infinite models, but these were disregarded from the start of this chapter.) The main other limitations of $\mathcal{SROIQ}$ are both shared by FOPL and irrelevant for REG. Let me give some examples. Consider REs such as

(a) Men who feed most (i.e., a majority of) dogs.
(b) Women who feed the same number of dogs and cats.

These REs cannot be expressed in $\mathcal{SROIQ}$, nor can they be expressed in FOPL (intuitively, one would need an infinitely long conjunction to express it). For REG, however, this does not matter, because if a referent is referable by one of the expressions above, then it is also expressible by a "less abstract" RE. Consider example (a), assuming there are exactly n dogs in the domain. Each of the men in the referent set (i.e., the men who feed a majority of dogs) feeds a specific number of dogs. Let n be the smallest number of dogs fed by any of the men in the referent set (so n exceeds half the total number of dogs), then GROWL will find the RE $Man \sqcap \geq nFeed.Dog$. This RE identifies the target set. The RE of example (b) is constructed in similar fashion. The situation is different of course if the exact number of dogs fed by each man cannot be inferred from the Knowledge Base.

1 No similar restriction hold regarding the disjunction of relations. For even though role disjunction, as in the quasi-formula $Man \sqcap \exists Adore \sqcup Criticize.Child$ ("the men who adore or criticize a child"), is not expressible, an equivalent concept can be expressed through disjunction of concepts, as in $Man \sqcap ((\exists Adore.Child) \sqcup (\exists Criticize.Child))$.

11.3 Even More Expressive Algorithms?

Although the "reading off" process employed by the Referability Theorem was not intended to produce a REG algorithm, and although full FOPL deduction does not have the efficient reasoning support that exists for SROIQ, the fact that "reading off" can be done constructively suggests the possibility of new REG algorithms. One straightforward algorithm would proceed as in Algorithm 20. The algorithm exploits the fact that the formula that is *read off* the model generated by the intended referent r must distinguish r from each distractor that it can be distinguished from.

Algorithm 20 A primitive, logically complete algorithm for referring to singular referents

Input: A domain of objects, with properties and 2-place relations defined on it. The domain contains a target referent r and a non-empty set of distractors.
Output: A distinguishing description of r if one exists. "Reading off" is defined in section 11.1.

1: Compute the submodel $M(r)$ generated by r
2: Read off a FOPL formula φ from $M(r)$
3: **if** r is the only x for which $\varphi(x)$ **then**
4: optimize φ, resulting in φ'
5: **return** φ'

Without optimization (line 4), this procedure would produce unnecessarily lengthy descriptions. Different approaches to optimization are possible (cf., section 8.5), including local optimization (cf., [Reiter, 1990b]), which removes conjuncts that are not necessary for producing a distinguishing description.[2]

In the absence of computationally efficient procedures, and in the absence of evidence for the linguistic reality of the REs that can be generated along these lines, we do not pursue this line of work any further here. It is time to abandon Part III of the book, which has targetted the generation of types of noun phrases that earlier algorithms were unable to generate (namely, logically complex REs and REs that contain a proper name). Leaving this part of the book

2 Note that φ can contain a wide range of formulas, including negated ones such as $\neg\exists x A(c, x)$. Hence, to test whether r is the only object for which φ is true (line 3), it does not suffice to proceed analogous to [Krahmer et al., 2003], testing whether $M(r)$ stands in the relation of subgraph isomorphism to any other parts of M.

behind us, we turn to a number of recent studies that are starting to challenge established conceptions of what reference is. Unlike the challenges faced in Part III, the difficulty of these further challenges will not lie in the types of referring expressions that are generated, but in the situation in which they are generated.

11.4 Summary of the Chapter

Chapters 8-10 have shown how the expressive power of REG algorithms can be extended beyond the classic REG algorithms. The present chapter has helped us understand this extended expressivity, by putting it in context, showing how these algorithms compare to what is possible in principle. The upshot of our discussion is as follows:

- We have discussed what it might mean for a REG algorithm to be logically complete in an absolute sense. The Referability Theorem proves that the resulting sense of logical completeness coincides with a notion of definability in First-Order Predicate Logic (FOPL) with equality. [Section 11.1]
- The SROIQ algorithm of the previous chapter, although it is more expressive than any other REG algorithm, is not logically complete in this sense, because it is possible for an entity to be non-referable by SROIQ even though a FOPL-based formula can identify the entity uniquely. [Section 11.2]
- The process of "reading off" a FOPL formula off the model generated by an intended referent could be used in principle to let REG algorithms achieve full logical completeness (in the above-defined sense). Here, however, we enter a territory where efficient generation may no longer be achievable. Likewise, we may be approaching a point where human speakers struggle to find distinguishing descriptions; so REG algorithms of this kind should be regarded with some hesitation. Earlier, in section 10.7, I have offered some reasons why these algorithms, and the REs they generate, are worth investigating nonetheless. [Section 11.3]

IV FOURTH PART: GENERALIZING REFERENCE GENERATION

Reference seemed such a simple idea: to refer is to anchor your utterances to "things", ensuring that people will know what you are talking about. As such, reference is an example of Information Sharing (sections 1.7 and 3.1): information already shared between speaker and hearer enables the speaker to pass on her privileged information to the speaker, so this information becomes shared as well. This is quite a simple idea indeed.

Yet this simple idea covers a large range of communicative situations. This Fourth Part of the book will discuss complications of Information Sharing, by focussing on situations that challenge existing approaches to REG. These challenges arise in situations that do not meet the presuppositions behind the classic REG task definition that were listed in section 4.4. To explain the challenges, I chose not to use formal definitions, but informal communication scenarios. As we shall see, each challenge has led to new REG algorithms that widen our understanding of reference.

The first challenge (chapter 12) arises in complex domains, where crucial information may only be obtained through some "fact finding". The second challenge (chapter 13) stems from situations in which the speaker does not know what the hearer knows. The third challenge (chapter 14) arises when the referent resists being singled out precisely, so REs can at best be approximative. The fourth challenge (chapter 15) arises when the speaker has something else in mind than the identification of the referent.

Various computational solutions to these four challenges will be discussed. Frequent echoes of the Battle of Balaclava will be heard, which featured in chapter 1, and whose fatally misunderstood reference to "the front" caused mayhem. These echoes will be especially loud when we discuss situations in which speakers struggle to fathom what hearers know (chapter 13), or where hearers have to dig deep to discover "the lay of the land" (chapter 12).

Because the problems discussed in this fourth Part of the book are difficult, and cognitive scientists have only recently started to investigate them, we shall paint the models that have been proposed with a broad brush: to expound on the details of an algorithm that may still be far from its final formulation would not be useful.[1]

1 The discussion of the First Challenge makes extensive use of [Paraboni et al., 2007] and [Paraboni and van Deemter, 2014], although the presentation of the JOVE algorithm is modified. My discussion of the Second and Fourth Challenge owes a debt to [Kutlak et al., 2011], [Kutlak et al., 2012], and [Kutlak et al., 2013]. The discussion of the Third Challenge elaborates on [Khan et al., 2006], [van Deemter, 2009a], and [van Deemter, 2009b].

12 *First Challenge:* Large Domains

In this book we have often presented domain knowledge in a table, neatly arranged to highlight the commonalities and differences between the entities in the domain. But real life is seldom so accommodating: in large domains we often have to work hard to find out how the facts line up.

The first challenge to existing models arises when a domain is large and complex, making it difficult for the recipient to grasp it completely. In these situations, which are highly relevant at a time when "big data" attracts broad attention, hearers may have to explore the domain, trying to find the referent: its properties may not yet be part of the information that the sender and the recipient share. Consider the following communication scenario.

The Book scenario. *Someone is reading a bulky textbook and has got to page 30, where it says "See picture 168". Because pictures are numbered throughout the book, this can refer to only one picture. However, this description would cause the reader unnecessary work. A more elaborate one, such as "See picture 168 in chapter 14" or "See picture 168, on page 420", would have saved her time.*

The challenge for REG is to find algorithms that produce elaborate REs when they are needed. This requires a departure from existing referential strategies. The text on page 30 of the book that is mentioned in the scenario identifies the picture uniquely, so why add more information? There are two possibilities: either the location of the picture is known to the reader, in which case there is no need to say more than "picture 168"; or it is not known to the reader, in which case it may seem useless to mention it. The idea that information not yet known to the hearer can help her find the referent is incompatible with the theory embodied in the algorithms discussed so far.

The amount of effort that a well-crafted RE could have saved here is somewhat limited in this case, because books are designed to be easy to browse. Moreover, readers may have seen only a handful of pictures on pages 1-30 and estimate that picture 168 is probably somewhere in the last quarter of the book, allowing them to speed up their search. To see that the problems can be worse, let us turn to a more challenging scenario:

The Direction Giving scenario. *Suppose Lewes Road is the longest street in Brighton. I invite you to come to number 968, the highest number in the street. You have never visited Brighton before and have no map. I tell you, speaking in a cafe in London, to "come visit me in Brighton tomorrow, at*

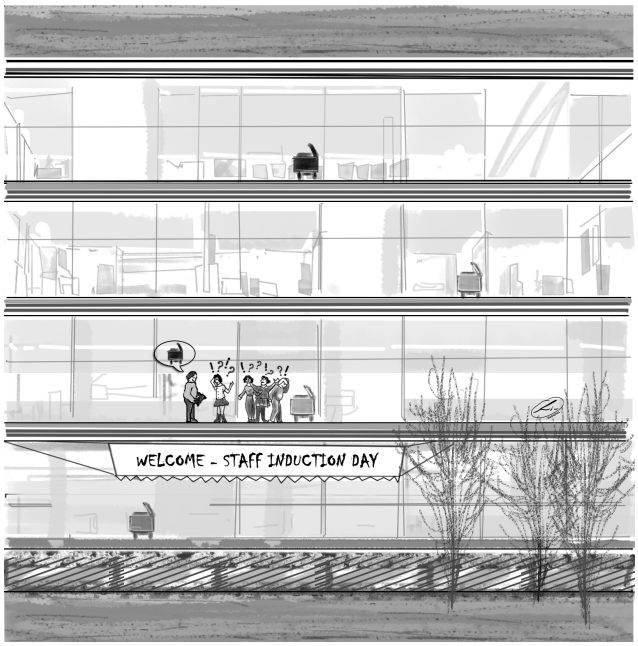


Figure 12.1
Lack of Orientation (LO): "Use the black copier. There's only one, you can't miss it."

number 968". Because other streets in Brighton, being shorter, only have numbers lower than 968, this is a distinguishing description. Yet it does not help you much to find the house. A description like "968 Lewes Road" ("... in the Moulsecoomb area, right at the end of the street") could have saved you a huge amount of time.

(Figure 12.1 shows a rather similar situation.) Ivandré Paraboni, Judith Masthoff, and I started by studying the Book scenario, conducting experiments and constructing algorithms [Paraboni et al., 2007], after which Paraboni and I turned to Direction Giving [Paraboni and van Deemter, 2014], focussing on spatial domains. Instead of testing participants' behaviour in the real world, we chose to focus on the interactive virtual environments provided by the GIVE evaluation challenge [Koller et al., 2010b].

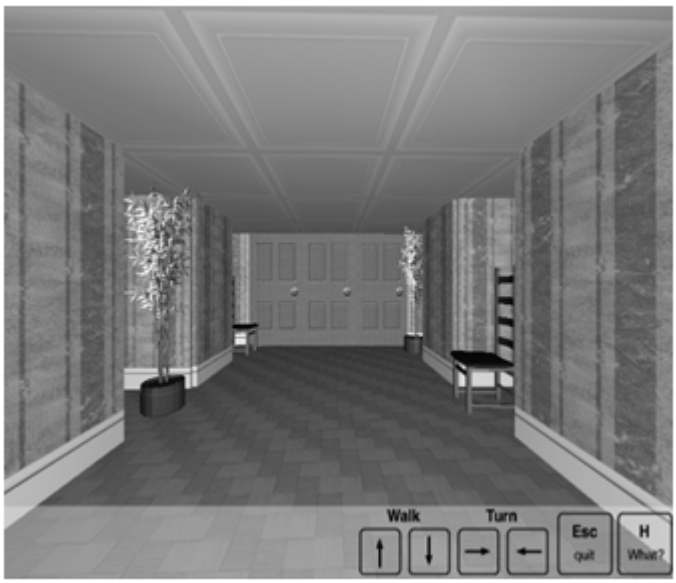


Figure 12.2

A domain as used in the GIVE *game (Striegnitz et. al., 2011), rendered here in black and white.*

A "world" of the GIVE game consists of a 3D virtual space containing rooms with doors, tables, and chairs (Figure 12.2), and with buttons in it that can be pushed. Virtual environments, although more complex than the domains studied in earlier chapters, lack many of the complexities of the real world but they allowed us to control the details of the environment precisely; they also allowed us to measure search effort in a variety of ways, by logging the time taken and the distance travelled by a subject who is searching for the referent.

Our work on the Book scenario had taught us that two types of referential situations are particularly problematic, which we termed Lack of Orientation (LO) and Dead End (DE). These terms turned out to apply straightforwardly to spatial domains, so let's introduce them. Note that some buttons are hidden behind the larger landmark objects (e.g., plants). It was therefore natural to refer to the less accessible object (i.e., the button) via the accessible landmark, thereby facilitating the hearer's search task:

(a) the blue button, behind the plant

Suppose there is only one blue button. Mentioning the plant is not logically necessary to permit identification, yet the simpler RE (b) may lead to what we call Lack of Orientation (LO), because the hearer has no clue as to the location of the referent:

(b) the blue button

As for DE, let us modify the scenario: this time there are several blue buttons, so the reference to the landmark (i.e., the plant) *is* necessary for disambiguation. Yet even (a) may cause problems if the hearer comes across *another plant* (with no blue button behind it) before reaching the intended plant: seeing this other plant, the reader may be puzzled not to find a blue button behind it. This is what we call Dead End (DE). Search would be facilitated if some additional, logically redundant information was added, as in:

(c) the blue button behind a plant, *on the right*.

Paraboni demonstrated that the descriptions generated by existing REG algorithms, which do not add logically redundant information in situations of this kind, slow down readers very considerably in terms of the time they require to find the referent, and in terms of the distance they travel (in digital space of course) during their search, when compared to more elaborate REs. To prevent this loss of time and energy, we proposed a procedure called Judicious Overspecification (JOVE), which enhances existing REG algorithms. The idea is to self-monitor the REs that a given REG algorithm produces – different REG algorithms can play this role – and to add information to the RE that is generated by this algorithm if the monitoring reveals the existence of an LO or DE problem. The term *over*-specification reflects that the REs in question contain properties that are not *logically* necessary for identifying the referent. There is nothing intrinsically wrong with over-specification: in fact, an overspecified RE may well be the shortest RE that any reasonable speaker would produce.

The pseudo-code of the JOVE algorithm (Algorithm 21 below) does not specify fully how an object is to be individuated: the choice of information to be used in overspecification is not spelled out completely; in the GIVE setting, a specific kind of location information ("on the left", "on the right", etc.) was used. This assumes that each object is located in one well-defined visual "context" that is small enough to be taken in visually by the player of the Direction Giving game, namely a room. Let's assume that the target referent is in the same room as the hearer (who begins his search in the position *start*, see Figure 12.3).

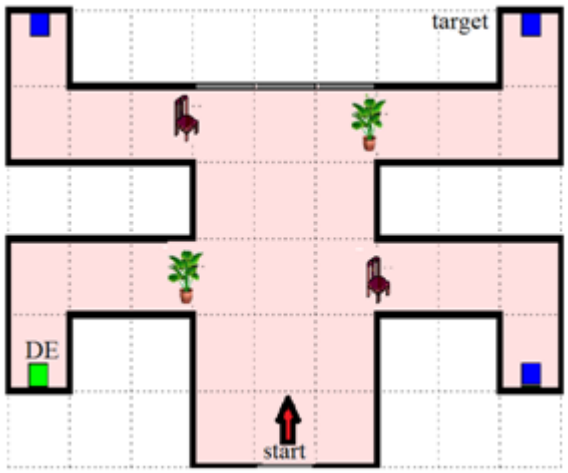


Figure 12.3
*Schematic view of a Dead End (*DE*) situation caused by the* RE *"the blue button behind the plant"*

JOVE assumes that a tentative description $\mathcal{D}$ for the referent r has been produced by an existing REG algorithm (e.g., sections 6.4 and 6.5). $\mathcal{D}$ may be simple or complex. It is possible that $\mathcal{D}$ describes r entirely by means of its own properties (e.g., r is the round blue button), without referring to any landmarks; in this case, JOVE focusses on r itself, trying to identify r. Alternatively, $\mathcal{D}$ may contain a landmark (e.g., the door in "... near the door"), in which case JOVE tries to identify this landmark. A landmark may be identified by means of another landmark, and so on.

If we simplify by abstracting away from the relations ("above", "near", etc.) between the various x_i and x_{i+1}, we can model a description of x_n as $\langle(x_1, P_1), (x_2, P_2)...(x_n, P_n)\rangle$, where x_1 is r itself, and each x_i is interpreted by means of the landmark x_{i+1}. For example, x_1 may be a button, P_1 may be the property of being a blue button, x_2 may be a plant, so P_2 is the property of being a plant. If this is the end of it, then $\mathcal{D} = \langle(x_1, P_1), (x_2, P_2)\rangle$. If the relation is "behind", then $\mathcal{D}$ encodes the RE "the blue button behind the plant".

The idea of the JOVE algorithm is to start from an initial description of this kind, to figure out how the player will search for the referent given this initial description, to detect any problems that may arise during this search, and to prevent these problems by adding information to the initial description. JOVE's

behaviour is summarized by the pseudo-code in Algorithm 21 below. The algorithm imagines that the hearer will go searching for the referent, following a path through the room. PATHS is the set of all search paths that the reader may choose, going from where he is, all the way up to x_n. PATHS is a *set* because the reader's search path is not fully known. Not every search path is equally plausible, and in [Paraboni and van Deemter, 2014] we have proposed and tested a simple model, called Nearest-First Search (NFS), of human searching. In what follows, we shall identify PATHS with the set of search paths permitted by NFS. Once PATHS is computed, JOVE monitors the RE, testing whether it may lead to LO or DE. If the reference leads to neither LO nor DE, then the RE is complete. If it does lead to LO or DE, then an additional property P is included in P_n, and this process of adding properties continues until the risk of LO/DE has been averted (see the While loop in lines 3-4 of Algorithm 21).

To see how NFS works, we return to Figure 12.3, divided into 5 contexts, as shown in Figure 12.4: the central area containing "start" and the landmark objects (i.e., chairs and plants); and four corridors containing one button each.

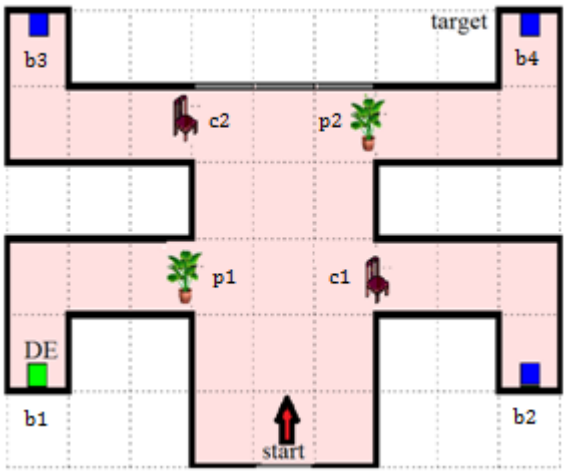


Figure 12.4
The domain of Figure 12.3. For ease of reference, unique names are added to domain objects.

Nearest-First Search (NFS) models the search by a reader for a referent, following the utterance of an RE:

1 The reader will exhaustively search for the target referent within the current visual context before considering any other visual context.

2 A decision to move to another visual context can never disobey the instruction provided by the RE (e.g., the reader cannot search on the right if the RE says "on the left"). If no instruction is provided, the hearer may choose to perform search in another context.

3 The object nearest the reader that matches the description is taken to be the intended target. If, later on, this interpretation turns out to be mistaken, the hearer resumes search as in clause 1. If two or more objects are equally distant from the hearer, the choice between them is random.

4 No object is inspected twice, and search stops when the goal has been reached.

Consider the search for "the blue button behind a plant" in the domain in Figure 12.4. The local context (i.e., the central area) is inspected first. Clause 1 of NFS dictates that within this area, the hearer searches exhaustively for the referent of "the plant". Then clause 2 dictates that this information be used to search for "the blue button" behind the corresponding plant.

Search may either end up in rapid success, or it may require a considerable number of steps, for instance if the hearer inspects p1, then b1, before reaching the correct target area behind p2. More precisely, non-problematic NSF-compatible paths to the target b4 are (c1,p2,b4) and (c1,c2,p2,b4). Additionally, a number of problematic paths (leading to DE when finding p1 and b1) are also compatible with NFS; examples include (p1,b1,p2,b4), (p1,b1,c1,p2,b4), (c1,p1,b1,p2,b4), and (c1,c2,p1,b1,p2,b4), among others. Thus, NFS defines a set of paths from start to target, each of which is considered a path that a person could conceivably follow on his search for the referent. The set PATHS on which the algorithm below rests contains precisely these paths.

The way in which JOVE checks the helpfulness of a RE is reminiscent of the classic REG algorithms of chapter 4, which test routinely whether a referent has been singled out by a given combination of properties. Some early dialogue systems such as HAM-ANS employed an *anticipation feedback loop*, which allowed the generator to monitor various aspects of its own output, including syntactic ambiguities and the clarity of its REs [Jameson, 1983] [Wahlster and Kobsa, 1989]. JOVE adds an empirically tested model of how readers search for a referent.

Going from experimental data to an algorithmic model invariably involves some extrapolation, because the data cannot cover all situations (e.g., rooms of all shapes). Based on our experiments with LO and DE, we proposed the following [Paraboni and van Deemter, 2014].

Algorithm 21 JOVE: Judicious Overspecification

Input: A spatial domain containing a hearer location l and another location where the target referent r is positioned. A set $\mathcal{P}$ of properties defined on the domain. In a description of the form $\mathcal{D} = \langle (x_1, P_1), (x_2, P_2) ... (x_n, P_n) \rangle$, where $n \geq 1$, the referent r equals x_1, and x_{i+1} is a landmark that helps to locate x_i (as in "the button behind the plant", where the button is x_i and the plant x_{i+1}).

Output: an RE $\mathcal{D}$, which may contain overspecification. The notion of a "problematic" path (and hence the choice of whether or not to overspecify) is explained below.

1: Let an existing REG algorithm produce an initial description $\mathcal{D} = \langle (x_1, P_1), (x_2, P_2) ... (x_n, P_n) \rangle$, where $n \geq 1$
2: PATHS := the set of NFS-compatible search paths from l to x_n
3: **while** some of the elements of PATHS are problematic **do**
4: Expand $\mathcal{D}$ by adding a suitable new location property to P_n
5: **return** $\mathcal{D}$

Suppose at some stage the algorithm is trying to describe an entity x_n. For example, x_n may be a landmark that helps to identify the intended referent x_{n-1}. Suppose the algorithm has ascribed the properties in the set P_n to x_n. Now a search path $O \in$ PATHS is problematic in two possible situations. A DE-type problem arises if O contains an entity of the same type as x_n (yet different from x_n) for which all properties in P_n hold true, and which may consequently be confused with x_n. The second type of problematic situation, associated with term LO, can be defined in different ways: we deviate here from the definition in [Paraboni and van Deemter, 2014] by saying that a search path has an LO-type problem if the path contains a pair (x_n,P_n) where x_n is not located in Context(x_{n-1}) yet there is no information in P_n that makes this clear. The idea is that x_n is too far away to get noticed by the hearer, and there is nothing in the description that points towards x_n. A LO-type problem arises in the domain depicted in Figure 12.4 if the starting instruction says "push the green button": there is only one green button, but it is in a different room, so JOVE considers it to be too far away to be easily found.

Example. Consider the situation in Figure 12.3. Initially, an identifying description is produced, for example, "the blue button behind a plant", where the button is x_1 and the plant is x_2 (and $n = 2$). The reference "a plant" to x_2 is submitted to JOVE, and tested for LO/DE. If the hearer finds the nearest plant first (as, in fact, the NFS search model of [Paraboni and van Deemter, 2014] predicts), then this is a DE-style obstacle (i.e., the short description "the blue button behind a plant" would cause DE). The fact that the wrong plant *may* be

encountered first (i.e., Paths contains a path in which this is the case) is sufficient reason for JOVE to amplify the initial Logical Form by adding information (e.g., "on the right"). The algorithm produces the non-minimal description "the blue button behind a plant, *on the right*". This expanded RE does not cause problems, so this overspecified RE is generated.

Are the scenarios discussed here really examples of reference, or is direction giving a different task (cf., [Dale et al., 2005])? This is a difficult question, because reference and direction giving are tied up closely with each other. Whenever a referent finds itself among a reasonably large number of distractors, REs need to contain information that guides hearers towards the referent (e.g., [Arts, 2004]). This is not only when the referent's properties are to be verified in physical space (i.e., observed), but equally when verification takes place in memory. Suppose, for example, I told you I heard a performance of music by a classical composer whose surname contains the letters "l" and "d". You might take some time to locate the referent in your memory; additional information (e.g., his nationality and the time in which he lived) will make it far easier for you to retrieve Haendel as a potential referent. Whether this is reference or Direction Giving is a moot question.

These scenarios do not deviate very much from the ones commonly studied in REG, yet they have shown that existing REG can be strikingly unhelpful. Doubtless, JOVE is only scratching the surface of the complications involved in them. The same journal issue in which [Paraboni and van Deemter, 2014] was published contains another account of the same problem, in which Garoufi and Koller use a transparent corpus (likewise based on the GIVE game) to train a *maximum entropy* model to allow a planning-based REG program to produce REs that minimize the time taken by hearers to find the referent [Garoufi and Koller, 2014]. Features include, for instance: the number of distractors in the room that are of the same type as the referent; the distance between the hearer and the referent; and whether the referent's colour is shared by one or more other objects in the domain. Other work based on Machine Learning uses Bayesian Classifiers to model speakers' choice of landmarks [Barclay, 2010].

Whether a rule-based approach or an approach based on Machine Learning works best is for further research to flesh out; in both cases, details can no doubt be improved. What is interesting from the present perspective is the need for checking whether a given RE is likely to cause readers difficulty. Both our

own and Garoufi and Koller's algorithm do this and might therefore be said to embody a simple Theory of Mind (cf., chapter 3, section 3.2).[1]

Psycholinguistic research (see section 3.2) has taught us not to assume too readily that speakers design their REs optimally for their audience. The work discussed above highlights a situation in which the lack of any audience design would cause disaster. The work reported in [Paraboni and van Deemter, 2014] – which focuses on the effects that REs have on recipients – does not tell us what speakers actually do, but earlier work contains a small study suggesting that human *authors* of written texts are good at this type of audience design when they are encouraged to think about the choice of RE in a text [Paraboni et al., 2007]. It may well be that human *speakers* (as opposed to authors) are frequently more egocentric, especially under time pressure and when referential success is not of crucial importance. When speakers do engage in the type of behaviour suggested by the JOVE model, they perform what is known in psycholinguistics as *self-monitoring* (e.g., [Levelt, 1989], chapter 12): they examine the output of the Conceptualization process (see section 1.4)), and if they observe that it falls short, they elaborate.

1 Compare [Koller et al., 2010a], which argues in favour of a "lightweight", behaviour-oriented model. See also [Engonopoulos et al., 2013].

13 *Second Challenge:* Breakdown of Common Knowledge

The previous chapter discussed situations in which the hearer is unable to identify the target until he has discovered some new facts about the domain. A different challenge to established accounts of reference and Information Sharing comes from situations in which the speaker is uncertain about the hearer's knowledge. Scenarios of this kind arise whenever we address an open-ended audience (i.e., when we publish something) and touch upon the philosophical problems discussed in section 2.2. Here is one such scenario:

> **The "Who is?" scenario.** *The publishing industry comes up with a new reading gadget. This gadget allows the reader of a document to select a proper name x that occurs in the document, asking "Who (or what) is x?" Its aim is to answer these questions in an optimally useful way. How should they be answered, assuming that space does not allow an entire Wikipedia page (or a similar extensive information source) to be displayed?*

The speaker does not know what the hearer knows, and this causes common knowledge (chapter 3, section 3.1) to break down. Breakdowns of common knowledge are a feature of life, because speakers can rarely be sure what their hearers know. Things could not have been more different in studies such as the TUNA experiment of chapter 5, where participants were invited to describe the referent that they saw on the screen in terms that the computer would understand. Although the computer did not know who the participants were, the information displayed was assumed to be shared between the computer and the participants. In fact, the furniture pictures in that experiment had been especially selected to ensure that their properties (type, colour, etc.) would be evident to any normally sighted viewer (see section 5.3). Thus it was reasonable to assume that all information about the scene was common knowledge.

The "Who is?" scenario was studied computationally by Roman Kutlák [Kutlak, 2014]. It resembles a scenario studied by Siddharthan, Nenkova, and McKeown, who likewise constructed algorithms that refer to people. However, these authors worked in a context of Text *Summarization*; when a generated summary had to refer to a person, it could "borrow" words and phrases from the texts that their algorithm was summarizing [Siddharthan et al., 2011]; moreover, much of their work focussed on the contextual appropriateness of a generated description (e.g., taking into account whether the person referred to had been mentioned earlier in the summary), which is not an issue in the "Who is?" scenario. Kutlák observed that the task of the gadget in this scenario fits Searle's definition (see section 1.1, and repeated here) precisely:

> "Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular I shall call a referring expression. Referring expressions point to particular things; they answer the questions Who?, What?, Which?" [Searle, 1969, pp. 26–27]

Yet this scenario poses some difficult challenges. Let us populate it. Suppose $x =$ the name "Jang Song Thaek", the name of an uncle of the North Korean president Kim Jong Un whose execution in December 2013 was widely reported around the world. The question "Who is?" is asked in the context of an Australian newspaper report in January 2014, which asserted that members of the man's direct family have now also been executed:

> News report of Jan 2014: "The North Korean dictator is reported to have put to death all of Jang Song Thaek's direct relatives."

Given the extensive media coverage of the events in December 2013, many readers will have known about the person executed, yet many will have forgotten the poor man's name. For them it makes sense to ask "Who is Jang Song Thaek?", and the reading gadget can help them by saying "This is the uncle of the North Korean president, who was executed in Dec 2013 after being denounced during a televized meeting. His execution was widely reported at the time." Other readers, however, may not know about the December events, in which case they are likely to be entirely unfamiliar with the referent. It is difficult to know what information would be useful to them.

Mindful of the difference between readers who do and readers who don't have previous exposure to the referent – *knowing* and *unknowing* readers, as we shall say – Kutlák focussed on the question of what information is most widely known about the referent. As for *knowing* readers, information that is widely known is relatively likely to be known by them as well; hence this information stands a good chance of triggering their memories of the referent. As for *unknowing* readers, information that is widely known is what these readers might be most interested in acquiring, because it helps to close the information gap between them and the people who do know the referent. The challenge, in both cases, is for the computer to find out what information is widely known.

Kutlák hypothesized that this problem may be solved if the internet is used as a window on hearers' knowledge. Many people read information on the world-wide web. If we do not know anything specific about a particular hearer's knowledge state, then this hearer might be modelled as having been exposed to a large but unknown fragment of the world-wide web. From this

perspective, if an item i of information occurs on the web more frequently than another item j, then the hearer is more likely to have been exposed to i than to j, so i is put earlier in the Preference Order than j.

Several versions of this idea were implemented and tested, the simplest of which is this: to find out how likely it is that an unknown reader knows that *Albert Einstein was a physicist*, search for documents on the world-wide web that contain the name "Albert Einstein" *and* the property "physicist". The more documents match this search, the more widely known this proposition is hypothesized to be. More sophisticated methods make use of Pointwise Mutual Information (PMI; [Fano, 1961]), which is based on a comparison between the joint probability of two events and the probability of co-occurrence of the two events if n and p are probabilistically independent of each other. In the present case, n is the occurrence of a particular name in a document and p is the occurrence of a particular property in the same document:[1]

$$PMI\ (n,p) = log\frac{P(n,p)}{P(n) * P(p)}. \tag{13.1}$$

If this was all there is to Kutlák's problem, and if the set of distractors is known (for example, by extracting a large set of famous people from the internet), then the monotonic approach to REG, introduced in section 4.5 might be resurrected to shape these ideas. In the pseudo-code of Algorithm 22, we leave out even more details than before.

Algorithm 22 REG algorithm based on an assessment of hearers' knowledge

Input: A domain of objects, containing a target referent r and a non-empty set M of distractors. A set $\mathcal{P}$ of properties of r. A metric (see text) that estimates how well known a property is in connection with r.
Output: A distinguishing description $\mathcal{D}$ of r if one exists.

1: Start out with an empty $\mathcal{D}$
2: **while** Not all distractors have been ruled out and $\mathcal{P} \neq$ empty **do**
3: Select a new property P from $\mathcal{P}$, *choosing the best-known one*
4: **if** P is false of some distractors **then**
5: Update $\mathcal{D}$, $\mathcal{P}$, and M

1 PMI itself did not have a good correlation with human judgments, but much better results were obtained when PMI was multiplied with $\sqrt{count(n,p)}$ to prevent pairs consisting of highly infrequent words that only occur together from achieving unreasonably high scores (e.g., if $P(n)$ is close to 0, then so is $P(n) * P(p)$). The idea of this correction stems from [Hodges et al., 1996].

In the "Who is?" scenario, however, it is unknown what distractors exist in the hearer's mind, and what properties he ascribes to each of them. The set of distractors is unknown; hence it is not possible to ascertain whether all distractors have been ruled out (clause 2) and whether P is false of some distractors (clause 4), so a different algorithm needs to be used, which does not rely on the distractor set.

Kutlák hypothesized that, once again, a crowd-sourced information resource like the internet can shed light, essentially by estimating the Discriminatory Power of an RE (as opposed to that of a single property). Once again, documents that express a certain conjunction of properties are retrieved from the world-wide web. One option is to linguistically "realize" this conjunction by expressing it as an English description, and to use this description to search the internet. If the proportion of documents returned that also contain the name of the referent is small, then the RE is considered not to be complete yet, in which case the next-best known property is added. For example, if the description of Einstein composed so far is "German-born physicist" and fewer than, say, 10% of the documents returned by the description contain the name "Albert Einstein", then the algorithm should continue to add properties to the Logical Form. The generation algorithm terminates when the number of documents retrieved that contain the name of the referent exceeds the threshold of 10%.

A number of variants of these ideas were first tested in pilot experiments. The two most promising ones, which were identical except for their termination heuristic, were then employed to generate expressions that refer to a number of famous people, based on the facts about these people represented in DBpedia, a huge database extracted from wikipedia [Bizer et al., 2009]. The information in DBpedia is extracted mostly from wikipedia *infoboxes*, which contain factual information such as persons birth date and birth place. DBpedia is a growing resource now containing structured information about 4 million entities, $600,000$ of which are people (as opposed to organizations, books, etc.). The output of algorithms was assessed via an experiment based on Mechanical Turk, in which participants were shown an English description and given the instructions in Figure 13.1. In response to the description, participants were asked to move a slider to the left or the right, depending on the extent of their agreement or disagreement with a given statement.[2]

2 A very similar use of this "magnitude estimation" method was reported in section 8.7.

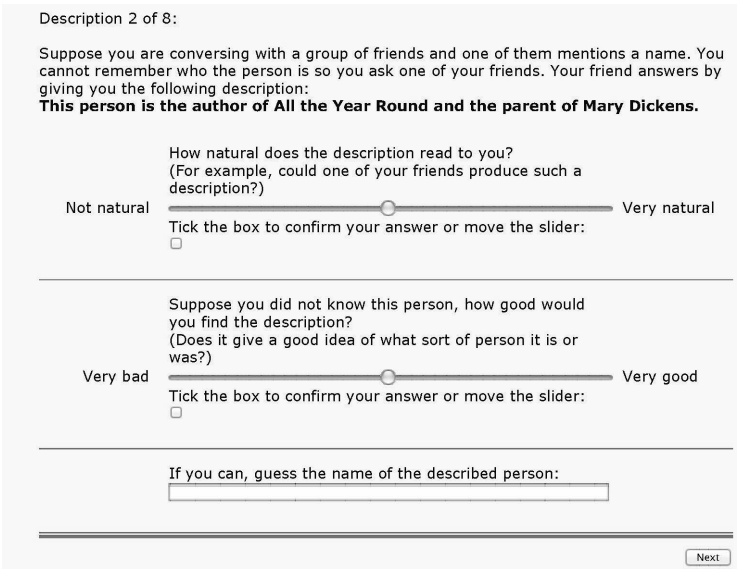


Figure 13.1

A part of the questionnaire employed in Kutlák's final evaluation experiment.

Evaluation was performed in terms of the number of people *correctly identified* by participants who were shown a number of descriptions that were generated, and in terms of the two other evaluation questions (Figure 13.1). Kutlák's algorithms were compared against a number of competitors, including a version of the Incremental Algorithm, and a set of specially crafted ("gold standard") descriptions produced by human authors. The results were encouraging. In terms of correct identification and quality (i.e., answers to the question *Suppose you did not know this person, how good would you find the description?*), both algorithms performed considerably better than the Incremental Algorithm. In terms of naturalness (i.e., participants' answer to the question *How natural does the description read to you?*) there was no statistically significant difference. Unsurprisingly, gold-standard descriptions composed by human authors outperformed their competitors in all respects [Kutlak et al., 2013]. Some examples of the descriptions compared can be found in the table below.

Kutlák's work has shown the radical implications of scenarios like the ones discussed in this chapter, and how the problems that are inherent in these scenarios can be addressed by using open, crowd-sourced, information

Referent	Algorithm	Description
Charles Darwin	Human	This person is considered the father of the modern theory of evolution due to his book On the Origin of Species.
	IA-optimal	This person was a British scientist who was popularised by Alvar Ellegard.
	Kutlák-1	This person was the author of On the Origin of Species, died in Downe, was known for evolution, natural selection, On the Origin of Species, and common descent, and named Notochthamalus.
	Kutlák-2	This person died in Downe, was known for On the Origin of Species and named Notochthamalus.
Audrey Hepburn	Human	This actress is famous for her performances in Breakfast at Tiffanys, My Fair Lady and Charade.
	IA-optimal	This person starred in the film Green Mansions and was a humanitarian.
	Kutlák-1	This person starred in Funny Face and lies buried in Vaud.
	Kutlák-2	This person starred in the film The Secret People and Funny Face and lies buried in Vaud.
Billie Holiday	Human	This person was a noted jazz singer, famous for her songs God Bless the Child, Lady Sings the Blues and Strange Fruit.
	IA-optimal	This person is a musician and is the author of A Mothers Gift.
	Kutlák-1	This person wrote Our Love Is Different, sang the songs Oh, Where Can You Be?; Lover Man; and Our Love Is Different; was also known as Lady Day; recorded The Lady Sings; and has homepage http...
	Kutlák-2	This person wrote Our Love Is Different, was also known as Lady Day, and has homepage http...

Table 13.1

Some descriptions generated by Kutlák's main algorithms (Kutlák-1 and Kutlák-2, which differ only in their termination procedures), the Incremental Algorithm with optimal Preference Order (IA-optimal), and a human-authored description (Human). Notochthalamus*, in descriptions of Darwin, is a species of barnacle.*

resources. Kutlák used the internet as the main source of information; and as his search queries were performed in English, the documents returned were also in English. These and similar methods could also be used with different languages, or focussing on other sources of knowledge, exploring other types of hearers' knowledge, and other communities' knowledge. These techniques are not only relevant for REG, but for all those situations in which one has to predict what information is likely to be available to a person about whom not a

lot is known. In particular, this is relevant to the Content Determination component of a Natural Language Generation system (section 1.4, see e.g., Figure 1.2). After all, in a reverse application of the ideas above, such systems should often *suppress* information that is unlikely to add to the reader's knowledge.

Many situations in real life combine features of the different scenarios discussed here. For example, it is worth reflecting on a scenario that shares some features with the "Who is?" scenario, and others with the Direction Giving scenario. The game theorist Barton Lipman described an imaginary game that features a speaker who needs an acquaintance to be picked up from an airport [Lipman, 2009]. Lipman, who was intrigued by the fact that human language makes such frequent use of vague expressions (e.g., "tall", "large", "grey") asked under what circumstances it would be beneficial, for the speaker and/or the hearer, if the speaker uses vaguely defined words (as opposed to exact measurements) as part of an RE. Here is a scenario that slightly different from the one discussed by Lipman, but directly inspired by it:

> **The Airport scenario.** *A speaker asks a hearer to go to the airport to pick up an acquaintance of the speaker. Unlike the speaker, the hearer does not know what the acquaintance looks like. There will be other people at the airport, but their number and features are not known in advance. What should the speaker say?*

Like in the Direction Giving scenario, the hearer will have to inspect the domain to find the facts. But unlike that scenario, the speaker cannot know what information the hearer will find there (e.g., how many distractors will there be?). As in the "Who is?" scenario, the speaker does not know who the distractors are (from the point of view of the hearer), yet the referential task is easier this time, because the speaker knows she should focus on visible properties (e.g., height, hair colour, etc.). However, she does not know what combination of properties suffices to identify the referent, hence she does not know how elaborate to be when she refers.

Lipman argued, using a Game-Theoretical perspective, that in situations of this kind, speakers who want to minimize the chance of misunderstandings should be as informative as possible. In the Lipman's original scenario this meant avoiding vagueness and rounding; in our own Airport scenario, it means conveying all those properties that the hearer will be able to check – a strategy that tends to lead to an extremely lengthy description, of course. Lipman noted rightly, however, that his arguments did not take into account that lengthy REs may be difficult to produce and to interpret, creating a potential trade-off

between clarity and effort [van Deemter, 2009a]. Further discussion of these issues would be out of place here, but it would be interesting to explore experimentally what is the optimal amount of information that the speaker should pack into her REs [Green and van Deemter, 2011] and how actual speakers perform such trade-offs.

The scenarios discussed so far in this fourth Part of the book are a reminder of the artificiality and simplicity of the referential situations that are usually studied in REG. Let us now turn to another way in which reference in daily life tends to differ from the situations that are typically studied because, in these situations it is impossible, or not practically feasible, to separate the target referent from all possible distractors.

14 *Third Challenge:* Approximate Reference

Most REG algorithms assume that identification of the referent is a requirement. Moreover, they operate on the basis that all distractors should be removed from the hearer's attention, and that one distractor cannot be more important than another. In many situations, however, these assumptions are not justified. Once again, let us reflect on some scenarios.

The Olive Oil scenario. *You are preparing a meal in a friend's house, and you wish to obtain, from your own kitchen, a bottle of Italian extra virgin olive oil. You phone home to ask your young son to bring it round for you. You know that also in your kitchen cupboard are some distractors: one bottle each of Spanish extra virgin olive oil, Italian non-extra virgin olive oil, cheap vegetable oil, linseed oil (for varnishing) and camphorated oil (which is medicinal). It is imperative that you do not get the linseed or camphorated oil, and preferable that you receive olive oil. The expression "Italian extra virgin olive oil" guarantees clarity but may overload your helper's abilities. A very short expression, "oil", is risky. Perhaps you should settle for the intermediate "olive oil".*

Writing about this scenario, Graeme Ritchie, Imtiaz Khan, and I once sketched an approach to REG that operates by searching for the RE that has the lowest *cost* [Khan et al., 2006]. Unlike the costs of section 6.5, this time, the cost of a Logical Form S is defined in terms of a combination of the length of an RE (the Brevity cost, $f_B(S)$, with B for Brevity) and the number of distractors that the RE fails to remove (the Clarity cost, $f_C(S)$), for example, Cost(S) = $f_B(S) + f_C(S)$. Search for the lowest-cost RE could be greedy, always selecting the property that reduces Cost the most, or it could work in such a way that a minimal-cost descriptions is always found. Clarity is essentially a form of utility, of course, hence Clarity cost is a form of negative utility.

A failure to remove linseed oil should bear a higher cost than a failure to remove Spanish extra virgin olive oil. Therefore, instead of a simple linear function of the size of the set of distractors that an RE fails to remove, there is a curve where the cost drops more steeply as the more undesirable distractors are excluded. For example, each object could be assigned a numerical rating of how undesirable it is, with the target having a score of zero. The brevity cost function $f_B(S)$ could still be a linear function [Khan et al., 2006].

An algorithm along these lines could sometimes terminate before all distractors have been removed. Consequently, if the generator opts for saying "bring me the olive oil", then despite the definite article, it is not uttering a distinguishing description, a bit as if it had said "(...) one of the bottles of olive oil".

Figure 14.1
The chef to his wife: "I dropped the fish. Can you clean the area to your right for me?"

This suggests once again (cf., section 2.4) that the borderline between definite and indefinite NPs can be blurred and that techniques developed for generating one type of NP can be useful for generating the other.

Approximate REs also come up when the aim of a description is to refer to a spatial region. [Khan et al., 2006] offered a variant of the following scenario as an example. As in the Olive Oil scenario, it is difficult to remove all distractors, but this time the difference in status between different distractors does not come to the generator's aid.

The Dirty Floor scenario. *Mr X has dropped a piece of fish on the floor, then removes it. He would now like Mrs X to wipe the area clean. The fish doesn't leave a visible stain, so he has to explain where it was. It appears that there is no such thing as a distinguishing description (except the insufficiently informative "where I dropped the fish"), although Mr X can arbitrarily increase precision by adding further information, for instance "near the table", "on your right", and so on.*

An ideal description would cover the dirty area and nothing more. On the other hand, a larger area will also do. The domain is defined as all conceivable sub-areas of the floor, so the target is one element of the domain (i.e., one sub-area). The function $f_C(S)$ should now assess the aptness of the denotation of any potential RE S ensuring that this denotation contains the target (i.e., the contaminated area), and that it does not contain too much else.[1] In many cases, an RE that denotes the referent and nothing else does not exist.

Computerized weather forecasts (e.g., [Turner et al., 2009]) resemble the Dirty Floor scenario. These forecasts can assess which roads are likely to be icy, and hence dangerous. Systems of this kind inform road gritters concerning the condition of roads, to help them decide which ones require treatment with salt to avoid traffic accidents. Thousands of roads can be dangerous on a given night, so it is often not feasible to say exactly which ones.

The Road Gritting scenario. *Suppose one weather warning says "Roads in the Highlands will be icy" whereas another says "Roads above 500 metres will be icy". The first may have 10 false positives (i.e., roads gritted unnecessarily) and 10 false negatives (i.e., dangerous roads not gritted); the second has 100 false positives and only 2 false negatives. Which of the two references to the intended geographical area is preferable?*

This resembles the Dirty Floor scenario except that the weather warning is based on sensor readings from a finite set of points on the map. But although this makes it possible in principle to list precisely which points are predicted to be icy, this would lead to enormously lengthy descriptions. Scenarios of this kind can be very complex, because one and the same area can be approximated in different ways. In an American context, for example, one might say "The American Midwest" (describing a given area as a whole), "The Great Lakes region and the Eastern Great Plains region" (splitting the same area in two), or one might list all the States that cover the area: "Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin" (splitting the area into 11 parts).[2]

1 An added complication is that expressions such as "near the edge of the table" and "on the left" have fuzzy borderlines, that is, areas of which it is unclear whether they are properly described by these expressions; compare chapter 9, where vague properties are discussed.

2 The problem of finding a suitable partitioning of a geographical region is precisely analogous to the problem of partitioning a set, which came up in sections 8.7 and 8.8 and is discussed more fully in [Gatt and van Deemter, 2007] and [Gatt, 2007]. For example, the question comes up when it is felicitous to describe an area by means of partially *overlapping* sub-regions.

Note that it matters which of these summaries is generated, because each summary will lead to a different set of roads being treated with salt (i.e., gritted). Gritting all the roads is not an option, because salt is bad for the environment, but failing to grip a slippery road can cause traffic accidents. Faced with these difficulties, Turner and colleagues decided to disallow false negatives (i.e., every road predicted to be icy must be reported as such). In other words, their generator would select a third weather warning, which covers the entire target area but may have numerous false positives (e.g., "roads outside coastal areas will be icy").

Turner's cautious approach may well have been justified given the potential gravity of traffic accidents, but if the method is to generalize to other situations, where the advantages of false positives and false negatives are more finely balanced, then a more flexible approach may be called for. One perspective on these matters would, once again, associate each utterance with a *utility*. To utter a sentence, after all, is to perform an action, and the choice between different actions is naturally thought of as governed by utility maximization, where utility is understood in the broadest possible sense, subsuming factors such as Clarity and Brevity. The idea, in other words, is to look at NLG as an area of applied decision theory.

Given a decision-theoretic framework of this kind, then, in the scenario above, one might reason that if a false positive has a negative utility (i.e., cost) of 0.1 and a false negative has a cost of 0.5, for example, then the first summary is preferable to the second. For the first summary only has a cost of $(10 * 0.1) + (10 * 0.5) = 6$, whereas the second has a cost of $(100 * 0.1) + (2 * 0.5) = 11$.

The analysis of language production as driven by the utility of utterances (see [van Deemter, 2009a]) appears to offer a reasonable answer to the problems presented by the Dirty Floor and Olive Oil scenarios (above), and it feels natural to people familiar with practical applications, where texts are generated for a concrete purpose. This approach is far from a *panacea*, however, and needs to be studied in more detail. The choice of the utility weights associated with the different kinds of error is crucial, of course, and can be tricky to justify, as is illustrated by the Road Gritting scenario. Also, although our emphasis on utility is only able to shed light on the question what REs are best for *hearers*, it is as yet unclear to what extent *speakers* are sensitive to this type of utility (cf., section 1.5, where two different perspectives on Natural Language Generation were contrasted).

15 *Fourth Challenge:* Going Beyond Identification

When we describe an entity, identification of that entity is not always our main aim. To gain an understanding of what other aims a description can have, and how these aims may be modelled computationally, we examine some scenarios in which identification of the referent is one factor alongside others.

Let us cast our minds back to the COCONUT experiment of section 6.6, in which pairs of participants buy furniture together. The dialogues between participants see them proposing a piece of furniture, persuading their partner to buy it, changing an earlier proposal, confirming an earlier proposal, and so on. The Intentional Influences model [Jordan, 2000c] contained rules that guarantee identification of the referent. Additionally, the model contains rules such as the following, that serve a different purpose:

> If the context is a *persuade* context, then select the properties that make the item a good solution to the problem.

This rule applies when there is a need to convince the hearer of the merits of a particular furniture item. These and other rules in the model show how the identification of a referent can go hand in hand with other communicative goals. Recent years have seen work on *Recommender Systems* that follows a very similar logic when a system figures out what to say about an item to a particular user [Carenini and Moore, 2001], [Tintarev and Masthoff, 2012]. Recently, Steven Knox, Erik Sanstedt, Paul Zah, and other undergraduate students at Aberdeen studied a similar type of situation, focussing on reference in a sales situation in which the preferences of the user are not known:

> **The Camera Adviser scenario.** *You are discussing cameras with a sales adviser. You are looking for a high-end compact camera and have discussed some options with the adviser, who has mentioned 5 different cameras to you. At a loss from so much choice, you ask "Which camera should I buy?" The answer should refer to a specific camera, in such a way that the choice for this camera is motivated.*

Common though this type of scenario is, existing REG algorithms are not well placed to handle it. For although the adviser is trying to identify a camera, identification is only a part of what she is trying to do when she says "The Nikon with the full-frame sensor": her aim is to let the most important features of the camera get across to you. If your interests, as a customer, are very well understood – you may be particularly interested in a camera's ability to take good pictures in low light, for example – then these should be taken into

account. But what if not enough is known about the customer's interests? Can we predict "interestingness" computationally?[1]

Predicting interestingness is made easier, in this scenario, by the fact that *quantitative* attributes, such as price, weight, and pixel count, are so pervasive in it. In chapter 9, where gradable properties were discussed, we saw how such attributes can be handled; the approach proposed in that chapter can give precedence to properties that express extreme values. Suppose, for example, the cameras range in price from 100 to 1000 dollars, with most cameras costing somewhere around 300; then the property of costing 289 dollar is less "interesting" than the property of costing, say, 100 dollar, even if the property of costing (precisely) 289 dollar has huge Discriminatory Power. Extremity can be measured statistically as a Z score (i.e., comparing a value's distance to the mean of the set with the standard deviation over the set); in a pilot study, this approach appeared to give plausible results in relation to the Camera Adviser scenario, producing REs that express numerical information that camera buyers might care about.

The idea of the interestingness of a property can be used as yet another way to populate the monotonic REG algorithm scheme of section 4.5, which we have seen a number of times now. Consider Algorithm 23 below, for example, which also takes on board the idea from section 13, that properties are added regardless of whether they remove any distractors (thus deviating from algorithms that we have seen). Depending on how interestingness is defined,

Algorithm 23 REG algorithm based on "interestingness"

Input: A domain of objects, containing a target referent r and a non-empty set M of distractors. A set $\mathcal{P}$ of properties of r. A metric (see text) that tells us how interesting a property is in connection with r.
Output: A distinguishing description $\mathcal{D}$ of r if one exists.

1: Start out with an empty $\mathcal{D}$
2: **while** $\mathcal{P} \neq$ empty **do**
3: Select a new property P from $\mathcal{P}$, *choosing the most interesting one*
4: **if** P is false of some distractors **then**
5: Update $\mathcal{D}$, $\mathcal{P}$, and M

1 The Aristotelian notion of an *essential* property comes to mind: a property that an entity cannot fail to have without becoming another entity. The camera I'm looking at could have been painted a different colour without becoming a different camera; but if its sensor was replaced by a different one, then it would no longer be the same. Thus, sensor type might be seen as an essential property, and colour as an inessential one, perhaps because sensor type explains more of the camera's other qualities (e.g., price). I do not know whether there is computational mileage in this idea.

it may or may not be possible to rank properties in terms of interestingness *in advance*: if interestingness is understood simply in terms of extreme values (see above), then it can be computed once and for all for a given r. If interestingness is understood in terms of how surprising a property is, however, then it should depend on the other properties assigned to r. For example, a price tag of 3000 dollars may be extreme for a camera, but if the camera is known to be a Leica, an expensive brand, it is no longer surprising and may therefore be less notable.

Unfortunately, the notion of interestingness outlined above is only applicable in relation to quantitative variables; in other cases it is meaningless. Can we predict how interesting a property is when the property is not quantitative? This is a difficult question, but luckily not one that only researchers in NLG have asked: it has been studied in data mining, for example, where researchers have designed a range of metrics for automatically discovering interesting patterns in data. As Roman Kutlák realized, some of these metrics may be of use in REG situations where extremity of values does not apply.

In their survey of interestingness metrics for data mining, Geng and Hamilton suggested that interestingness may be determined by a variety of factors, including novelty, generality, reliability, and "actionability" (i.e., the extent to which the information in question is able to guide the hearer's actions), each of which one may attempt to measure. One of the most important factors is what the authors call surprisingness or unexpectedness [Geng and Hamilton, 2006], defined as patterns that contradict existing knowledge or expectations. In a study discussed in chapter 13, Kutlák applied this idea to REG by hypothesizing that unexpected, and thus interesting, properties of a person are properties that are rare for a person to have. For example, being awarded the Nobel prize is unexpected because only a handful of people receive this prize. By contrast, having a mother is rather expected and consequently it is not worth mentioning; having a famous mother is, once again, interesting.

Formalizations of unexpectedness tend to have conditional probability at their core. For example, when we compute the extent to which x's having won the Nobel Prize is unexpected, we want to know how likely it is for someone to win this prize. But what information should we assume as given? The fact that x is a physicist? The fact that x is female? At this stage, these are open questions. Moreover, the conditional probability of a fact may not tell the whole story, and other statistical constructs may be relevant. For example, being born on the 12th of March has low probability, yet someone's being born on that date is not very unexpected or newsworthy. Perhaps this is because in the case

of someone's birthday, where (information-theoretic) *entropy* is high, we do not build up an expectation, hence the 12th of March is not unexpected. But in a case where entropy is lower, as in the date when someone starts their PhD (which is often around the start of the academic year), the 12th of March might be unexpected. At the time of writing, these potential avenues for formalizing the notion of unexpectedness – and hence interestingness – are only starting to be explored. They will not be discussed further here, but see [Kutlak, 2014].

In the scenario above, identification is no longer the be-all and end-all of reference, but merely one factor among others. To show that "identification of the referent" can be pushed even further into the background, let's consider yet another scenario. Recent years have seen the emergence of computer systems that produce textual descriptions of photographs of everyday scenes. We focus here on the MIDGE system, which extends systems discussed in [Li et al., 2011] and [Yang et al., 2011], automaticaly generating descriptions of photographs gathered from Flickr; an even more recent example, based on the development of a huge corpus of human-annotated pictures, is [Kazemzadeh et al., 2014]. The authors of [Mitchell et al., 2012] describe MIDGE as "generating a natural sounding description of a photograph from computer vision detection" [Mitchell et al., 2012]. We take MIDGE to be operating in the following scenario:

> **The Flickr scenario.** *A person is shown a photograph of a scene that he or she may not have seen before. The person is asked to describe what's in the photograph, using just one* NP. *The instructions ask her to* "describe it as you might describe it to a blind person who wants to know what the photograph depicts. Please avoid any meta-information (e.g., whether it's a nice picture, or where the picture was taken)*"*.

MIDGE takes as input a set of entity detections, along with some of their properties and some spatial relations between them, as recognized by a computer vision system. This input is combined with statistical information extracted from a corpus of 700,000 images in which each picture is coupled with a human-generated verbal description, which is often a (definite or indefinite) noun phrase. The objects and substances recognized are associated with nouns, which are used as the seed of the generation process, during which descriptions are partly "hallucinated" on the basis of corpus-based probabilities.

For instance, two objects may be recognized next to each other, and these may be associated with the words "duck" and "pond". As a first step towards

combining these nouns into a textual description of the image, they are ordered from left to right, based on which order is most frequent in the corpus; for example, "duck" should occur before "pond". Each of these nouns gathers a set of small syntactic trees around it, each of which is consistent with the computer vision output and reflects the syntactic contexts in which the noun appears frequently in the corpus; for "duck" this might include a syntax tree for "duck [next to]", "duck near", "duck in"; for "pond" this might include syntax trees for "in [a pond]", "[next to] [a pond]", "near [the pond]", *etcetera*. These local trees are then combined into grammatically correct combinations, such as a syntax tree for the phrases "[a duck] [in [a pond]]", "[a duck] [near [a pond]]", "[a duck] [[next to] [a pond]]", etc. The final, complete description is produced using further corpus-based methods, for example, based on the frequency of the 3-grams that make up each description.

There may seem to be something rash in "hallucinating" so much of the content of these descriptions, but the limitations of computer vision make this, for the moment, the best way to make an educated guess at what may be going on in a picture, using a philosophy not dissimilar to the one employed in chapter 13. Evaluation of the descriptions produced by MIDGE suggests that the system does a reasonable job even though, as one might expect, human-produced descriptions received much higher ratings in terms of grammaticality, correctness, and humanlikeness. MIDGE-generated NPs include "People at a wooden table", "A brown cow", and "Boats under the sky" (all of which were broadly correct as descriptions of the pictures they describe), but also "the sky" (as a description of a picture that shows a horse in a meadow, but no sky). Leaving the technical performance of the system apart, what stands out from our viewpoint is the communicative task on which MIDGE puts the spotlight: what does it mean to "describe" an entity or a set of entities, in a situation where singling it out from a finite set of distractor entities is not the point (or not the main point) of the description?

MIDGE differs from the reference tasks that are the core of this book. Yet some insights from the Camera Adviser scenario and the "Who is?" scenario of chapter 13 extend to the Flickr scenario, despite the fact that the noun phrases occurring in the latter are often indefinite (using an indefinite article or a bare plural). First, if it is possible to identify the referent, then it helps to do this (if the photograph depicts the Eiffel Tower, it would be strange not to say so). Second, if you see something remarkable in the picture – a striking landscape, an extremely tall building, or a man biting a dog – then this is worth saying as well. Interestingness, in other words, seems as important in the Flickr scenario

as in the other two. What this suggests is that interestingness is not limited to descriptions and other NPs: it governs all our choices concerning the propositions that we decide to put into words. As we saw in connection with previous scenarios, ideas which govern reference production are of considerable relevance to other areas of language and communication as well.

Moving away even further from the classic reference task, one can think of scenarios such as the following:

> **The Unnamed Company scenario.** *Two people in my research group have founded a company, Arria* NLG. *When colleagues ask who they visited last week, they may say "(we visited)* a large oil company whom we are hoping to get as a customer*". Their description is not designed to allow us to identify the referent, but to give us an insight into the reasons for their visit, and the type of business their company are interested in. In a* variant *scenario, there are business reasons why the two are not able to disclose the identity of their prospective customer.*

The speakers' task starts with a specific company. As before, their task is to provide information about this entity, mindful of the interests of the hearer. The difference is that they do not try to identify the referent, because they consider this uninteresting to their audience. In the variant scenario, the plan is *not* to permit identification of the referent, leaving something like interestingness as the main reason for including a property.[2] One could imagine more extreme scenarios, in which the aim is to deceive the hearer concerning the identity of a referent. What these examples demonstrate is that a speaker can have very different aims with their descriptions.

In the Unnamed Company scenario, it would have been possible (although possibly ill advised) for the speakers to identify the referent for the hearer. In other cases identification is not even possible:

> **The Musical Chairs scenario.** *You and your sister are standing outside a room where a children's party is in progress. Inside the room, the music stops, signalling that a game of musical chairs has just concluded. You tell your sister: "one player is without a chair". She nodds and you continue: "Go inside the room and offer a lemonade to the player without a chair".*

2 Note that this makes the scenario similar to what happens in the computational task of data anonymization (mentioned in section 1.2), where information is filtered to ensure that referents cannot be identified.

Given the rules of the game, some players must be without a chair, and because every player tries to grab a chair, there is unlikely to be more than one of them. Consequently, the Russellian statement (cf., section 2.4), that exactly one player is without a chair, holds true. This is what makes your utterance (to your sister) make sense. It is hard to say which of the NPs in the scenario *refer*: some theoreticians would call "the player without a chair" an attributive description (section 2.6). Be this as it may, the rule-based approach of section 10.6 was designed to tackle situations of exactly this kind, enabling a generator to refer to entities that are only *inferred* to exist.

We are seeing that the purpose of a description can vary considerably: it can be to identify the entity precisely (as in the classic REG problem), or to identify it approximately (as in the Road Gritting scenario of chapter 14), or to explain what's striking about it (as in the Flickr scenario), or combinations of the above (as in the Camera Adviser scenario). Considerations such as the following are applicable to a range of descriptions:

The extent to which a property can help to recommend an object

The extent to which a quantitative property is extreme

The extent to which a property is surprising or unexpected

Earlier chapters have shown that the Maxim of Relation (which covers the notion of relevance) is less well understood than the other Grician Maxims. But relevance is implicated in many of the considerations above. For example, a feature that recommends the referent in one context may not recommend it in another; a feature that is surprising in one context may not be surprising in another. A proper understanding of the issues discussed in the present section therefore seems to be a crucial step towards a full understanding of reference, and other areas of communication as well. For once again, these are issues that affect language production in all its facets: they apply not only when we refer, but whenever we communicate.

Summary of Part IV: Complexities of Information Sharing

In Part I, we defined *Information Sharing* as the process in which a sender exploits information that she shares with the recipient, to anchor her utterances to objects (section 1.7). This idea has remained central to our account. Parts II and III demonstrated that the nature of the information exploited can vary considerably: from a simple database of atomic facts, on the one hand, to a set of complex axioms (section 10.6), on the other. Now, in Part IV, we have seen that the idea of Information Sharing hides some important distinctions, for example between:

a. Situations in which the speaker has incomplete information concerning the recipient's knowledge. The Airport Scenario and the "Who is?" scenario were examples of this problem.

b. Situations in which the recipient lacks some crucial information about the domain that she can nonetheless recover by searching. Examples include the Book scenario and the Direction Giving scenario.

c. Situations in which there does not exist a distinguishing description, or none whose length is acceptable. The Dirty Floor scenario was an example of the former and the Road Gritting scenario an example of the latter.

d. Situations in which approximate identification suffices. In the Olive Oil scenario, some distractors need to be removed but others are less crucial; in the Dirty Floor and Road Gritting scenarios, all distractors are equally important but the task allows a limited number of false positives and/or false negatives.

Reference – or a task that is very similar to reference – can even take place in communicative settings where identification of the referent does not lie at the heart of the expression, including:

e. Situations in which (precise or approximate) identification plays a role without being the main aim. Often when we refer, we have an ulterior motive: to ask attention for something, to recommend the referent, or to warn against it. This happens, for example, in the Camera Adviser scenario.

f. Situations in which the aim of a description is to express what is interesting (e.g., because it is perceptually salient, or practically useful) about an entity. The Flickr scenario exemplifies this category.

g. Situations in which the aim of an NP is not to refer but to express a quantified statement, for example that one (or at least one) x has the property y. Examples include the Unnamed Company scenario and the Musical Chairs scenario.

It is easy to think of variants of these scenarios. For example, *collaboration* between the speaker and the hearer (discussed in sections 1.6 and 3.6) could be added to a number of these scenarios, and this would alter the dynamics of Information Sharing considerably.

V EPILOGUE

16 Epilogue

How can one gain an understanding of a complex cognitive phenomenon such as reference production? This book suggests that understanding requires collaboration between academic disciplines, with an interplay among theory, empirical experimentation, formal analysis, and computational modelling.

We started, in Part I, with an overview of the issues raised by philosophers and psycholinguists. In Part II we saw how computational linguists implement algorithms that produce REs from non-linguistic input, and how these algorithms are found and tested in increasingly sophisticated ways, based on experiments with human speakers and hearers. Part III showed how, in recent years, REG algorithms have started to address a much wider range of referring expressions (REs), enabling these algorithms to refer to sets, to make use of gradable properties and a variety of quantifiers. In Part IV we looked at a range of problematic scenarios, which put our understanding of reference and Information Sharing to the test and which led to new algorithms. As the reader of section 1 may recall, a number of these challenges are implicated in the disastrous misunderstanding that caused the defeat of the British army at Balaclava: Lord Raglan, in saying "advance rapidly to the front" was trying to refer to a geographical area that was not precisely delineated (just like in the Road Gritting scenario of section 14). He misjudged his hearer's view of the terrain (as may happen in the "Who Is?" scenario of section 13) and failed to make his utterance safer by adding more information (unlike the JOVE algorithm in the Direction Giving scenario of section 12). He could have added "on the Causeway Heights", for example, and this would have prevented the misunderstanding.

It is time to take stock. We start by putting REG into the context of Computational Cognitive Modelling (section 16.1); this area of research is usually seen as unrelated to REG, but I shall argue that a different view of REG is possible and, in fact, increasingly plausible. Next, we shall briefly revisit the Gricean Maxims – principles that guide a lot of research throughout Linguistic Pragmatics – in light of the insights embodied by REG algorithms and the principle of Intrinsic Preference, which underlies the Incremental Algorithm and its descendants (section 16.2). We conclude by discussing the future of research on reference production (section 16.3).

16.1 REG Algorithms as Cognitive Models

Historically, REG algorithms are rooted in practical applications, as we have seen. This book, however, has emphasized a different perspective, which views these algorithms as models – imperfect models, but models nonetheless – of reference production. It is interesting then to compare REG with other computational models of human abilities, looking for similarities and differences. Effectively, this means that REG will be viewed as an area of Computational Cognitive Modelling. As the starting point for this comparison I will use a summary of Computational Cognitive Modelling from the Introduction to the Handbook of Computational Psychology, where some existing classifications of the field are discussed [Sun, 2008]. I have extracted a number dimensions of variation from Sun's discussion and applied these to REG.

(1) First, Computational Cognitive Models differ according to their *aim*. Most commonly associated with Computational Cognitive Modelling are models that aim to simulate (i.e., mimic) human behaviour. Other models, however, have a different aim. For example, they may aim to produce output that is practically useful, possibly surpassing human performance in a particular area. An area where this distinction is often discussed is logical reasoning, with some models focussing on logically valid reasoning (the classic domain of Formal Logic), and others on actual human reasoning with all its peculiarities and deficiencies [Bringsjord, 2008].

One can see in table 16.1 that most REG models have tried to mimic human speakers (although a minority have focussed on the production of REs that allow hearers to find the referent quickly and reliably). Why is this? Given its roots in practical applications, a focus on utility for hearers might have been more apt. The reason, I think, is that computational linguists are used to evaluating their output by comparing with a corpus – the BLEU metric used in Machine Translation is one of the best known examples – so this is the method that REG researchers knew best. The method had to be adapted (cf., chapter 5), but once suitable transparent corpora were available, there was no problem in principle with applying it to REG. Only gradually did researchers start to realize that evaluation in terms of effectiveness (i.e., utility for hearers) makes at least as much practical sense as humanlikeness. The fact that REG evaluations started in terms of humanlikeness might be seen as a happy accident, because the idea of simulating speakers might not have been investigated in such depth otherwise.

(2) A second dimension discussed by Sun concerns the *granularity* of the input and output of a model. Focusing on the output, for example, a Computational Cognitive Model of social behaviour might focus solely on the number of Twitter feeds produced by an individual or, in a more fine-grained analysis, it may try to capture their topic as well. Similarly, a model of speech might focus merely on the articulation of words, or it might capture intonation patterns as well. In REG, a large number of studies have focussed specifically on the semantic content of generated REs, which is known as the problem of Content Determination (see section 1.4); only a minority of computational studies (discussed in chapter 8) have looked at Linguistic Realization as well (see table 16.1). So, once again, a distinction originally invented in connection with other Cognitive Models is applicable to REG.

(3) A distinction that is even more interesting in connection with REG is the one between *process* and *product* models [Vicente and Wang, 1998] [Sun, 2008]. Product models, also known as blackbox models or input-output theories, formalize the relation between the inputs to a system (e.g., a domain and an intended referent) and the outputs that it generates (e.g., an RE), without making any claims about the manner in which that mapping comes about. Some researchers feel that product models should not be regarded as cognitive models, because they do not tell us much about the manner in which the human mind works. In the opinion of these researchers, process models – which aim to model the manner in which people perform a cognitive task – are the only properly *cognitive* models. Others, however, beg to differ, arguing that the processing details of a model are often highly speculative, in which case it may be more realistic to regard the relations between inputs and outputs as the thing that a Computational Cognitive Model makes claims about.

To view an algorithm as a product model may be counter-intuitive, but when evaluation studies test a REG algorithm by comparing it to a corpus of human-produced REs (see e.g., chapter 5), they do look at the relation between inputs and outputs only; in other words: they test the quality of the algorithm as a product model. Looking at the field of computational REG as a whole, it is clear that most studies – such as the TUNA Evaluation Campaigns of chapter 5 – have treated REG algorithms as product models. At the time of writing, very few studies examine the processing implications of an entire REG algorithm; an example of a study where this does happen is a recent investigation that asks whether speakers' speech onset times follow the pattern that one should expect if speakers followed the Incremental Algorithm [Gatt et al., 2012]. Other studies are harder to place along this dimension: where a computational study is

informed by classic hypothesis testing, for instance, it might be viewed as involving process models, to the extent that the study addresses a principle (e.g., the principle of Intrinsic Preference) that plays a role in REG.

(4) Another of Sun's distinctions is the one between models of *individual* agents and models of *groups* of agents. It is interesting to see how this distinction plays out in the study of reference. At first sight, the vast majority of work on REG focusses on individual agents, because so far only a handful of computational studies have focussed on interaction (see Table 16.1). However, groups of agents can not only be studied in terms of the interactions between agents but also in terms of the differences between them. The former category was discussed in section 3.6; the latter category in section 6.1 (including the work of Viethen and Dale), and in sections 6.2 and 6.3 (including the PRO algorithm).

Sun discusses two other dimensions of variation that are of interest to us, namely, **(5)** the amount of *algorithmic detail* that is offered, and **(6)** the extent to which a model is inspired by the *physiology* of the human brain (see table 16.1). We have discussed the former at the start of section 10.5; we shall turn to the latter in section 16.3, focussing particularly on models inspired by specific evidence about language production rather than by general (and sometimes rather loose) considerations of cognitive-neurological architecture.

Looking at the table as a whole, it is evident that the mainstream of REG does not always match the mainstream of Computational Cognitive Modelling. For example, REG today is characterized by an emphasis on product models, as opposed to process models. One wonders whether it may be time for researchers in REG to ask more systematically than before how the working of their algorithms – as opposed to its mapping between inputs and outputs – might be tested, for instance by means of eye-tracking.

REG is not commonly seen as a part of Computational Cognitive Modelling, although our discussion suggests that it can be seen that way. If REG researchers were to enter the mainstream of Computational Cognitive Modelling, then this could have interesting consequences. For example, the interaction with learning, with short-term memory, and with other features of human cognition could take center-stage, perhaps as part of a more generic cognitive architecture, such as ACT-R (see e.g., [van Rij et al., 2013] for an application of ACT-R to reference comprehension) or Soar [Laird, 2012].

	Mainstream	Exceptions	Comments
1. **Aim**	Mimic REs that were elicited from human speakers	Produce useful REs [Paraboni et al., 2007] [Turner et al., 2009] [Koller et al., 2010b] [Garoufi and Koller, 2014]	This distinction was only made systematically after 2000
2. **Granularity**	Only perform Content Determination	Produce complete NPs [Stone and Webber, 1998] [Krahmer and Theune, 1998] [Khan, 2013]	
3. **Model the product or the process?**	Model only the *product* of human language production	Model the language production process itself [Gatt et al., 2012]	This distinction is usually left implicit
4. **Individuals or groups?**	Model individual behaviour	Model collaboration [Heeman and Hirst, 1995] [Garoufi and Koller, 2014]. Model group variation [Viethen and Dale, 2010] [van Gompel et al., 2012]	
5. **Algorithmic detail**	Has been implemented	Has not been implemented [Jordan, 2000a] [Ren et al., 2010]	Many algorithms have been implemented in multiple ways
6. **Physiological basis**			Not used in computational REG yet

Table 16.1

A classification of Computational Models of Referring, broadly inspired by the discussion of Computational Cognitive Models in [Sun, 2008]. A classification of non-computational work on reference production would show a different pattern, with more attention being paid to the production process, to social interaction, and to neuroscientific evidence.

16.2 The Gricean Maxims and the Principle of Intrinsic Preference

In chapter 3, we identified a number of factors that affect reference production, calling them Truthfulness, Discriminatory Power, Relevance, Clarity, and Intrinsic Preference. All except the last one derive from the Gricean Maxims [Grice, 1975], so let us revisit these Maxims briefly in light of our findings.

Quality: *"Do not say what you believe to be false. Do not say that for which you lack adequate evidence."* Herb Clark's work has modified these principles in such a way that the common ground shared by the speaker and hearer is taken into account (section 3.1), but this is not the end of the story. For not only have we observed that speakers' ability to meet these requirements is limited (section 3.2), we have also seen that situations arise in which it is not feasible for speakers to limit their REs to information that is in common ground (chapter 13): speakers, in these situations, have to be gamblers.

Relation: *"Be relevant."* Despite Kronfeld's insistence (see our section 4.2) on relevance, REG has long struggled to make algorithmic sense of this idea. From where we stand, one can see the beginnings of a solution. Pam Jordan, in her Intentional Influences model, demonstrated how the context of a dialogue can make properties salient that would normally be overlooked (section 6.6). Ross Turner explored the idea that an irrelevant RE is one that suggests an incorrect explanation for an event (section 3.3; see also chapter 14), for instance as when a weather forecast says "strong winds are expected in rural areas" (instead of, for example, "... in coastal areas"). Furthermore, we have seen in chapter 15 that statistical analysis can help to predict which properties may be surprising or interesting; this is another dimension of *relevance*, relating to the degree to which a property is relevant to the hearer's values and decisions. Finally, Albert Gatt has demonstrated, working on plural REs (section 8.7), how complex REs can be made lexically coherent; I have argued that this idea may also help ensure the relevance of other words in a text (section 8.7). This plurality of accounts suggests that relevance may not be one issue but many. Information may be relevant for different reasons, and our explanations for why an RE (or a part of an RE) is relevant should reflect this.

Manner: *"Avoid obscurity of expression. Avoid ambiguity. Be brief. Be orderly."* In the realm of REG, some of these issues can be covered by other Maxims: the injunction to be brief may be covered by the Maxim of Quantity, for instance (see below). Perhaps the most central of the four injunctions

is the one about avoiding ambiguity. Imtiaz Khan found that, although ambiguity is undesirable, it is productive to allow ambiguity under certain conditions (section 8.8). Like Ivandré Paraboni before him (see chapter 12), Khan found that clarity needs to be traded off against brevity. In Paraboni's case, brevity needed to be weighed against the amount of search effort inflicted on the hearer; in Khan's case, against the likelihood of misunderstandings. Grice, of course, was well aware that his Maxims could clash, so it is possible to see in these results a sharpening, rather than a refutation, of his ideas.

Quantity: *"Make your contribution as informative as is required. not make your contribution more informative than is required."* As chapter 4 shows, much of the history of REG is a series of qualifications of this Maxim: early investigators such as Winograd and Appelt discovered that straightforward interpretations of Quantity are at odds with the facts of human language production. Dale and Reiter asked attention for what we called the principle of Intrinsic Preference, culminating in their Incremental Algorithm.

Having revisited the Maxims, it is time to discuss Intrinsic Preference. We have seen that the Gricean Maxims alone do not offer a complete explanation of the patterns observed in reference production. Most researchers believe that something like a fixed Preference Order has to be taken into account as well, because some types of information are simply more "important" than others, for reasons unrelated to their Discriminatory Power or practical interest. Models of referring ignore these issues at their peril (see e.g., sections 3.5 and 6.2). Yet fixed Preference Orders should be taken with a pinch of salt. This is not only because of the logical and algorithmic problems that arise when the values of an attribute have overlapping extensions (section 4.7), but for empirical reasons as well. Let me explain.

Much of the attraction of the idea of a fixed Preference Order rests on the assumption that the values of an attribute have much in common, and that they must all be preferred to the same degree. For example, the task of finding a good Preference Order (as in the TUNA experiment of chapter 4) becomes much harder if different colours, different sizes, and so on, can be preferred to different degrees. The idea of an attribute as a cluster of similar properties has some initial plausibility, not least in terms of how the different values of an attribute are treated by the human information processing system. However, we have seen in sections 3.4, 3.5, and 6.3 that the degree of preference associated with an attribute can depend on many factors: a statement like "colour is

preferred over size" is at best a rule of thumb; at a deeper level, there are other issues at work, which cut across Attributes.

Other considerations point in the same direction, suggesting that the values of a given attribute may differ sharply in terms of their degree of preference. For example, blue and red are values of the same attribute, yet the colour of a blue strawberry is more likely to be noted than that of a red one, because the prototypical strawberry is red. The fact that not all values of an attribute are equally important is even clearer in the case of quantitatively valued attributes, such as distance, weight, and price: if you bought a cadillac for only 500 dollars, wouldn't this make you more likely to mention the price than if you had paid a more normal (i.e., higher) price? Consequently, the *extremity* of a property should be taken into account when determining whether the property is worth mentioning (chapter 15). This idea generalizes Hermann and Deutsch's old principle that if two objects differ along two dimensions in opposite directions, then the dimension that represents the largest of the two differences tends to be expressed in reference (section 3.4).

It might be added that the idea of incremental Content Determination does not sit easily with the order in which words are realized (in English and many other languages), as discussed in connection with gradable adjectives in section 9.5. This confirms that, despite recent progress, one should be extremely cautious regarding current REG algorithms as models of the reference production *process*: at best, the algorithms offer good models of the relation between inputs and outputs of human language production; the components and the timeline of the process are an entirely different matter.

Finally, all the evidence for Intrinsic Preference comes from a tiny set of attributes, namely, TYPE, COLOUR and SIZE. It is plausible that the TYPE of an object contributes disproportionately to the *Gestalt* that the object creates in the human mind. Similarly, COLOUR is perceived in different ways from other attributes, and often (though not always) with great vividness (section 3.4). So if experiments demonstrate that TYPE and COLOUR occur frequently in human-produced REs, then this should of course be taken into account by computational models. But to jump to the conclusion that *all* attributes can be linearly ordered in terms of their degree of preference, and that this plays a key role in human production of REs, would be unwarranted. Perhaps TYPE and COLOUR are exceptions that need to be treated in special ways, whereas other attributes are governed by Discriminatory Power. The exceptional status of the TYPE attribute is widely accepted. It is possible that COLOUR deserves special treatment too.

In short, Incremental Content Determination has to be regarded with caution; this conclusion is even more inevitable in connection with relational REs (section 6.4) and reference to sets (chapter 8), where incrementally runs into additional problems, as we have seen. The future belongs to a more flexible approach modelled loosely on the Greedy Algorithm. At the core of this approach lies a *monotonic* approach to the selection of properties first introduced in section 4.5, which might be generalized along the lines of Algorithm 24, and augmented along the lines of the PRO algorithm of section 6.3 and the JOVE algorithm of chapter 12 to make sure there is a place for overspecification. Needless to say, the pseudo-code below is a mere caricature of an algorithm, from which some crucial details are missing. The pseudo-code focusses on solutions to the classic REG problem, leaving the logical complexities of Part III (relational descriptions, reference to sets, the use of gradable properties, and the use of negations and disjunctions) aside.

Algorithm 24 The shape of a future solution to the classic REG problem?

Input: A domain of objects, containing a target referent r and a non-empty set M of distractors. A set $\mathcal{P}$ of properties of r.
Output: A distinguishing description $\mathcal{D}$ of r if one exists. $\mathcal{D}$ is chosen probabilistically, based an a variety of factors (see text).

```
Start out with an empty D
while Not enough distractors have been ruled out and P ≠ empty do
    Probabilistically select a new property p from P.
    if P is false of sufficiently many distractors then
        Update D, P, and M
while D makes r difficult to find and P ≠ empty do
    Add a suitable property from P to D
```

Factors influencing the selection of a new property have been discussed in the last few chapters, with Intrinsic Preference, Discriminatory Power, knowledge status, extremity, and interestingness all playing a role. How these factors combine is a question that is likely to occupy researchers for a long time, but our discussion in section 6.3 suggests that, if human reference production is to be modelled, the choice of a new property has to be nondeterministic, with different REs being produced on different occasions. Given that factors such as knowledge status and interestingness apply to many aspects of communication, this work my well prove to have practical relevance to Text Summarization, Recommender Systems, and the general problem of Content Determination in Natural Language Generation.

16.3 Future Research: The Way Ahead

We have argued in section 1.1 that the problem of modelling reference production is NLG complete, in the sense that solving REG would mean solving all of NLG. This makes the modelling of reference production a huge task, so it is important to prioritize the open problems in this area. Based on what we have seen in earlier chapters, let me offer some suggestions.

Investigate logically complex REs and proper names. I have emphasized that REs come in many shapes, and that the logical complexity found among REs poses difficult logical and computational challenges. Yet a comparison between Parts II and III of the book reveals a stark asymmetry: the empirical side of the classic REG problem (studied in Part II) has been investigated extensively: the main algorithms have been tested thoroughly with human participants, and many relevant hypotheses have been tested in psycholinguistic experiments, but the same is not true for proper names and complex REs (studied in Part III): although we saw a range of algorithms for the generation of logically complex REs, and a number of experimental studies underpinning some aspects of these algorithms, they have yet to be scrutinized systematically. A similar gap exists in psychology, where few experiments so far have focussed on complex REs. Until complex REs have been studied thoroughly, our knowledge of REG remains painfully incomplete.

Study axiom-based REG. The classic definition of the reference task (section 4.3) is limited to logically simple REs. The task of generating logically complex REs led us to use modern Knowledge Representation formalisms such as Description Logic (chapter 10). But Knowledge Representation not only allows us to construct complex REs; it also allows us to utilize generic and incomplete knowledge, which are widely recognized as natural components of human knowledge. In section 10.6, we have started to explore the implications of this view, which can enhance our understanding of *attributive* descriptions, a type of description much studied by philosophers of language. An axiom-based approach to REG, which makes full use of TBox as well as ABox axioms, goes against the grain of current trends in Computational Linguistics – which warn justifiably against hand-crafted rules – but could revolutionize REG and other areas of Computational Linguistics, aided by modern techniques for automated reasoning and years of experience with empirical methods.

Attributive descriptions and proper names are prominent topics of discussion in theoretical linguistics, philosophy of language, and the formal semantics of natural language. I have argued that it is time to study these topics computationally as well, and I have offered suggestions for how this may be done. As argued in chapter 2, some other theoretical issues might benefit from a computational approach as well; examples include misdescriptions (including metonymical and Donnellan-style descriptions) and the types of descriptions studied by Russell and Strawson (section 2.6).

Study complex communication scenarios. The study of complex scenarios in Part IV revealed that REG changes when Information Sharing is compromised. Far from being exceptional, the challenges discussed in Part IV are common in daily life. For example, speakers routinely refer despite only having incomplete information about hearers' knowledge (chapter 13), and in situations where the hearer needs to examine the domain before being able to resolve the RE. Furthermore, speakers often have to make do with approximate descriptions. So in these respects as well, the classic REG task definition of section 4.3 amounts to a stark simplification. Challenges such as the ones discussed in Part IV offer rich opportunities for further study.

Study variations between and within speakers. Given the considerable variations that exist in the REs produced by different speakers (and by one speaker on different occasions), a model of reference production overlooks something important unless it takes these variations into account. To exemplify this direction of travel, section 6.3 discussed the PRO algorithm, which produces, for each input, not just one RE but a probability distribution on a set of RE types. It will be interesting to see similar methods being applied to complex scenarios. Also of great interest is the kind of speaker modelling pursued by [Viethen and Dale, 2010] and others (section 6.1), which can be extended to study connections between personality characteristics and reference production. It seems likely, for example, that some speakers are better at handling epistemic mismatches than others; scenarios of this kind may pose particular challenges to people on the autistic spectrum, for instance [Arnold et al., 2009], [Wicklund, 2012], [Nadig et al., 2015].

Enhance the study of large and naturalistic domains. The research community knows much about the classic REG problem, except in connection with large and naturalistic domains. Psychological and computational work on reference has focussed on small and simple domains, involving only a handful of

objects, and compromizing ecological validity in various ways (section 3.7). Exceptions include investigations of the effect of clutter on reference production [Paraboni et al., 2007, Coco and Keller, 2009, Koolen et al., 2013a, Koolen et al., 2013b, Clarke et al., 2013, Paraboni and van Deemter, 2014]. Also exceptional for its attention to larger domains is Gorniak and Roy's work on the interpretation of REs: these authors describe a computer program that interprets REs elicited from speakers [Gorniak and Roy, 2004]; REs refer in a domain of up to 30 cones on a computer screen; domains are chosen in such a way that information about location is crucial, as in "The green one that's closest to us in the front". All these studies confirm that existing algorithms fall well short of capturing how speakers behave – and what hearers require – in large domains, and suggest that computational models of reference production need to scale up.

To indicate why reference to real-life objects can cause trouble for REG algorithms, let me list some of the things that Margaret Mitchell found when she studied her arts-and-crafts domain (see section 9.6). First, the notion of the TYPE of an object, which plays an important role in many REG algorithms, is problematic. For example, speakers referred to a heart-shaped object by calling it *the heart*; does this make HEART the TYPE of the object?[1] Second, when an object possesses a property that is prototypical for objects of a given type, then this makes it less likely that this property is mentioned ([Mitchell et al., 2013b]; compare section 3.6, and see [Westerbeek et al., 2015] for further developments).

Finally, and connected with the previous point, target referents are frequently described by *analogy* to something else. This phenomenon had been noticed in the psycholinguistic literature, as when the RE "the violinist" refers to a shape that reminds the speaker of a violinist [Clark and Wilkes-Gibbs, 1986a]. REG algorithms could emulate this type of behaviour by computing analogies in advance of the reference task. Focussing on the example above, pictures of some typical violinists could be stored in memory. Now if a Computer Vision program computes, for a given referent r, that r's similarity to one of the typical violinists is high, then this makes the property of "being similar to a violinist" true of r; when this property is added to the RE, all the domain objects that are dissimilar to the typical

1 A similar example is due to Richard Power (personal communication): reportedly, the word "palomino" denotes a horse whose coat is brown and whose mane and tail are white. Does this make palomino a TYPE? Does it denote a COLOUR?

violinists can be removed. Computing the required similarities in a way that matches human judgments is a challenging problem whose solution would not only help REG but contribute hugely to our insight in analogy, which is becoming an important theme in Artificial Intelligence and Cognitive Science [Hofstadter and Sander, 2013].

Focus on hearers as well as speakers. This book has paid considerable attention to experimental studies of hearers.[2] Nonetheless, the bulk of empirical work has so far emphasized emulation of speakers, rather than benefits for readers. The relative scarcity of work that investigates REG from a viewpoint of benefits for hearers (see section 16.1, first row of Table 16.1) is unfortunate, especially from a practical point of view. For, arguably, it is as important to know how to generate understandable REs as it is to know what REs human speakers tend to produce (cf., section 1.5). Recent years have seen links between computational REG and psychologists studying reference production; what is needed now is the establishment of similar links between computational REG and psychologists studying the *comprehension* of REs.

Aim for psychological and neurological realism. Much of this book is a plea for the idea that psychological models of language production should take the shape of algorithms. Yet few existing algorithms can claim to be good models of the human production *process* (see section 16.1, third row of Table 16.1).

When the generation process takes centre-stage, neuro-scientific methods are bound to come to the fore. Among the most established methods in neuro-cognition are methods based on measuring event-related potentials on the scalp (ERP); these can detect the time path of a mental task; another method, based on functional Magnetic Resonance Imaging (fMRI) scans, is more suitable for assessing the locations in the brain that are implicated in a task. Since about 2000, reference *comprehension* has been recognized as an important area of neuroscience [van Nieuwland et al., 2007], [Nieuwland and van Berkum, 2008], [Engelhardt et al., 2011], allegedly with its own distinctive neurological footprint, known as *Nref* (where the first character stands for *Negative* voltage and the remainder for *reference*). The Nref phenomenon has been confirmed using both ERP and fMRI.

2 This was especially true for Parts II and III, where we discussed the avoidance of ambiguity (section 8.8) and the design of coherent references to sets (section 8.7), and where we studied the production of REs under epistemically problematic situations (chapters 12, 13).

Neuro-cognitive methods can be aimed at the hearer, offering new ways of testing claims to the effect that one REG algorithm makes life easier for comprehenders than another (see e.g., the algorithms discussed in chapters 12 and 13). Moreover, they have the potential to offer insight into the human production process beyond what other methods, such as self-paced reading and eye-tracking, can teach us. If a reliable way can be found to measure ERP during language production, for example, this could give us a measure of the difficulty of a given reference task, whose outcomes could then be compared to the runtime of a REG algorithm (cf., [Gatt et al., 2012] for a study using speech onset times); note that a new focus on the production process itself could "rehabilitate" computational tractability as an important tool in the modelling of reference production (cf., section 4.8, which argued against the importance of computational tractability in this area of research). Furthermore, by telling us what areas of the brain are active during a particular production task, fMRI studies might tell us to what extent REG algorithms are on the right track when they posit problem solving activities such as "checking whether all distractors have been removed" and "adding information to help the reader", which play a role in some algorithms.

Neuro-cognitive methods are not easily applied to naturalistic language production, however. ERP is difficult to apply because the muscles involved in human speech articulation can easily affect ERP signals; fMRI is difficult to apply as well, because brain scanning does not permit participants to speak and act naturally. Various solutions to these problems could be explored, for example by letting participants in the experiment think of an utterance without actually producing (i.e., saying) it; moreover, initial studies of reference production based on Transcranial Magnetic Stimulation (TMS) of brain parts suggest that other methods might be able to circumvent the above-mentioned problems entirely (e.g., [Nozari et al., 2014]). It can be argued that, in the psychology of language, neuro-cognitive methods have yet to live up to their promise. Be that as it may, it now seems likely that neuro-cognitive studies of language production will soon fulfil their promise, not only by inspiring neurologically motivated computational frameworks (e.g., connectionism [Thomas and McClelland, 2008]), but by motivating the type of models that have formed the subject matter of this book.

These speculations conclude our exploration of reference as the *Drosophila melanogaster* (i.e., the fruit fly) of language, as we called it playfully in the Preface: a humble subject that has nonetheless attracted substantial interest from many corners of Cognitive Science, ranging from linguistics and the philosophy of language (chapter 2) to Artificial Intelligence, and from computational logic (e.g., chapter 10) to psychology (chapter 3) and, most recently, neuroscience.

As with the real fruit fly, scientists have both practical and theoretical reasons for studying reference. For example, like the fruit fly, reference is ubiquitous and relatively straightforward to experiment with. The fruit fly is often thought to be one of the simplest animals whose biology involves the same mechanisms (e.g., in terms of its genetics, physiology, metabolism, and life cycle) that are found in higher animals. Similarly, reference may be one of the simplest speech acts that involve the same mechanisms that are found in fully fledged communication, for instance in terms of Information Sharing, in terms of the relation between form and content, in terms of the role of logic and reasoning, and in terms of the variations that are observed both between and within speakers.

Clearly, many questions about reference and referring have yet to be solved. On the other hand, I hope to have shown that referring *is* a phenomenon that the Cognitive Sciences are starting to get quite a firm grasp of. The analogy with the fruit fly suggests that it may be time to investigate, more systematically than before, to what extent the insights gained in the study of reference production carry over to other phenomena in language and communication.

Frequently Occurring Terms and Abbreviations

Audience Design. The idea that human speakers "design" their utterances to make them optimally useful to hearers. The extent to which, and the conditions under which, speakers perform audience design is a much investigated issue in psycholinguistics. Related to the *Egocentricity debate*.

Common Ground. See *Shared Information*.

Common Knowledge. See *Shared Information*.

Conceptualization. The process whereby human speakers decide what information to express in their utterance (or in their *Referring Expression*). Conceptualization is the psychological counterpart to *Content Determination*.

Content Determination (CD). The process whereby a computer program determines the information content of a piece of text to be generated. CD is the computational counterpart to *Conceptualization*. This book mostly uses the term CD in a specialized sense, focussing on the informational content of a *Referring Expression*.

Definite Description. Definite Descriptions are *Noun Phrases* of the form "the so-and-so", or ones equivalent to these, such as genitives (e.g., "John's father", the father of John).

Denotation. The denotation of an English expression is the set of things that correspond with it in reality or in a model. Analogously, we speak of the denotation of a property. We use the term as synonymous to *extension*. The denotation of a is often written as $[\![a]\!]$.

Description Logic (DL). A family of Knowledge Representation formalisms closely related to decidable fragments of Predicate Logic.

Discriminatory Power (DP). The DP of a word or a property measures its capacity to remove *distractors*. DP is usualy formalized as the number of distractors removed, as a proportion of all the distractors present in a given situation (section 3.3).

Distinguishing Description. A *Referring Expression* that leaves no doubt about its intended referent. Like an RE, a distinguishing description can be a Noun Phrase or a Logical Form.

Distractor. A distractor is something we do *not* intend to refer to. In simple cases, where the intended referent is a single thing, a distractor is anything (in the model) except the intended referent. In pseudo-code, the referent is often represented by the symbol *r*.

Egocentricity debate. A topical debate in psychology that focusses on the question under what conditions adult people know, and realise, what other people know. Closely related to questions about *Audience Design*.

Extension. See *Denotation*.

Information Sharing. The process of turning privileged information into *Shared Information*. Usually, Information Sharing makes use of knowledge that is already shared between speaker and hearer. Part IV of the book discusses difficulties that can arise in Information Sharing.

Intensional. A function or process is intensional (with an "s") if it is able to distinguish between words or properties that have the same *extension*.

Knowledge Base. A computational device for storing knowledge (i.e., a glorified database). Most Knowledge Bases in this book store information that is in *Shared Information*. We use a variety of formalisms for representing knowledge, including *Description Logic*.

Logical Form. An expression of a formal language, designed to capture the information content of an RE. In its simplest form, the Logical Form is a set of properties, interpreted as logically conjoined. The more complex Logical Forms in Part III of the book use a greater variety of logical operators, such as negation, disjunction, and quantifiers.

Natural Language Generation (NLG). The use of computer programs for the generation of text in a human language. Generation is the computational counterpart of human Language Production.

Noun Phrase (NP). A syntactic category. For example, any expression that can be the subject of a sentence is an NP. *Referring Expressions* often take the form of a Noun Phrase.

Referring Expression (RE). By first approximation, REs are Noun Phrases whose aim is to single out a referent (see section 2.1). By extension, we also use the term to denote the information content of an RE, that is, a Logical Form.

Referring Expressions Generation (REG). The computational aspect of this book: the study of algorithms that model the production of referring expressions. Also known as Generation of Referring Expressions (GRE).

Shared Information. (also: Common Knowledge; Common Ground). Informally speaking, this is information *publicly* shared by a group of people. If information is shared, then each member of the group knows that it is shared.

Bibliography

[Abbott, 2010] Abbott, B. (2010). *Reference*. Oxford University Press, Oxford.

[Aloni, 2002] Aloni, M. (2002). Questions under cover. In Barker-Plummer, D., Beaver, D., van Benthem, J., and de Luzio, P. S., editors, *Words, Proofs, and Diagrams*. CSLI Publications, Dordrecht.

[Alshawi, 1987] Alshawi, H. (1987). *Memory and Context for Language Interpretation*. Cambridge University Press, New York.

[Anderson et al., 1991] Anderson, A. A., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34:351–366.

[Appelt, 1985a] Appelt, D. (1985a). Planning english referring expressions. *Artificial Intelligence*, 26(1):1–33.

[Appelt, 1985b] Appelt, D. (1985b). Some pragmatic issues in the planning of definite and indefinite noun phrases. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*.

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

[Areces et al., 2008] Areces, C., Koller, A., and Striegnitz., K. (2008). Referring expressions as formulas of description logic. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG-08)*.

[Ariel, 1988] Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24:67–87.

[Arnold et al., 2009] Arnold, J., Bennetto, L., and Diehl, J. (2009). Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition*, 110(2):131–146.

[Arnold, 2008] Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, 23:495–527.

[Arts, 2004] Arts, A. (2004). *Overspecification in Instructive Texts*. Unpublished PhD thesis, Tilburg University.

[Asher and Bonevac, 2005] Asher, N. and Bonevac, D. (2005). Free choice permission is strong permission. *Synthese*, 145(3):303–323.

[Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (2003). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK.

[Bach, 1987] Bach, K. (1987). *Thought and Reference*. Clarendon Press, Oxford.

[Barclay, 2010] Barclay, M. (2010). *Reference Object Choice in Spatial Language: Machine and Human Models*. Universiy of Exeter, Exeter.

[Bard, 2007] Bard, E. (2007). Let's you do that: sharing the cognitive burdens of dialogue. *Journal of Memory and Language*, 57:616–641.

[Bard and Aylett, 2004] Bard, E. and Aylett, M. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. In Trueswell, J. C. and Tanenhaus, M. K., editors, *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*. MIT Press, Cambridge, Mass.

[Barr et al., 2013] Barr, D., van Deemter, K., and Fernández, R. (2013). Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of ENLG 2013, The 14th European Workshop on Natural Language Generation*, pages 157–161, Sofia, Bulgaria.

[Barwise and Cooper, 1981] Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

[Bateman and Zock, 2002] Bateman, J. and Zock, M. (2002). *The John Bateman and Michael Zock List of Natural Language Generation Systems*. World-Wide Web at http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/.

[Beaver, 1997] Beaver, D. (1997). Presupposition. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 939–1009. North Holland.

[Belke and Meyer, 2002] Belke, E. and Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

[Belz, 2007] Belz, A. (2007). Automatic generation of weather forecast texts using comprehensive probabilistic generation space models. *Natural Language Engineering*, 14(4):431–455.

[Belz and Gatt, 2007] Belz, A. and Gatt, A. (2007). The attribute selection for gre challenge: Overview and evaluation results. In *Proceedings of UCNLG+MT: Language Generation and Machine Translation*.

[Belz and Gatt, 2008] Belz, A. and Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*.

[Belz et al., 2008] Belz, A., Kow, E., Viethen, J., and Gatt, A. (2008). The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 183–191.

[Belz et al., 2010] Belz, A., Kow, E., Viethen, J., and Gatt, A. (2010). Generating referring expressions in context: The GREC task evaluation challenges. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, pages 294–327. Springer Verlag, Berlin.

[Benson, 2012] Benson, T. (2012). *Principles of Health Interoperability HL7 and SNOMED*. Springer, London.

[Beun and Cremers, 1998] Beun, R. J. and Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6(1/2):121–152.

[Bierwisch, 1989] Bierwisch, M. (1989). The semantics of gradation. In Bierwisch, M. and Lang, E., editors, *Dimensional Adjectives*, pages 71–261. Springer Verlag, Berlin.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

[Blutner, 2000] Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17(3):189–216.

[Bod, 1998] Bod, R. (1998). *An Experience-Based Theory of Language*. CSLI Publications, Stanford, CA.

[Boër and Lycan, 1986] Boër, S. E. and Lycan, W. G. (1986). *Knowing Who*. MIT Press, Cambridge, Mass.

[Bohnet, 2008] Bohnet, B. (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 (INLG-08)*.

[Bott and Noveck, 2004] Bott, L. and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language*, 51:437–457.

[Brachman and Schmolze, 1985] Brachman, R. J. and Schmolze, J. G. (1985). An overview of the KL–ONE knowledge representation system. *Cognitive Science*, 9(2):171–216.

[Breheny, 2008] Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, 25(2):93–139.

[Breheny et al., 2006] Breheny, R., Katsos, N., and Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100:434–463.

[Brennan and Clark, 1996] Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

[Bringsjord, 2008] Bringsjord, S. (2008). Declarative/logic-based cognitive modeling. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press.

[Brown-Schmidt, 2009] Brown-Schmidt, S. (2009). Partner–specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, 61:171–190.

[Brown-Schmidt and E.Konopka, 2011] Brown-Schmidt, S. and E.Konopka, A. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, 2:302–326.

[Brown-Schmidt and Tanenhaus, 2004] Brown-Schmidt, S. and Tanenhaus, M. (2004). Priming and alignment: Mechanism or consequence? *Behavioral and Brain Sciences*, 27:193–194.

[Brown-Schmidt and Tanenhaus, 2006] Brown-Schmidt, S. and Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54:592–609.

[Bruner, 1983] Bruner, J. (1983). *Child's Talk: Learning to Use Language*. Norton, New York.

[Brunswik, 1943] Brunswik, E. (1943). Organismic achievement and environmental probability. *The Psychological Review*, 50:255–272.

[Bruza et al., 2009] Bruza, P., Busemeyer, J., and Gabora, L. (2009). Introduction to the special issue on quantum cognition. *Journal of Mathematical Psychology*, 53:303–305.

[Burge, 1973] Burge, T. (1973). Reference and proper names. *The Journal of Philosophy*, 70:425–439.

[Callaway and Lester, 2002] Callaway, C. and Lester, J. (2002). Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, Philadelphia.

[Campana et al., 2004] Campana, E., Tanenhaus, M., Allen, J., and Remington, R. (2004). Evaluating cognitive load in spoken language interfaces using a dual-task paradigm. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP'04)*.

[Carenini and Moore, 2001] Carenini, G. and Moore, D. (2001). An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 1307–1314, Seattle, WA.

[Carletta and Mellish, 1996a] Carletta, J. and Mellish, C. (1996a). Risk–taking and recovery in task–oriented dialogue. *Journal of Pragmatics*, 26:71–107.

[Carletta and Mellish, 1996b] Carletta, J. and Mellish, C. (1996b). Risk-taking and recovery in task-oriented dialogues. *Journal of Pragmatics*, 26:71–107.

[Carnap, 1947] Carnap, R. (1947). *Meaning and Necessity*. University of Chicago Press, Chicago.

[Cassell et al., 2000] Cassell, J., Sullivan, J., Prevost, S., and (Eds.), E. C. (2000). *Embodied Conversational Agents*. MIT Press, Cambridge, Mass.

[Chafe, 1980] Chafe, W. L., editor (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

[Chantree et al., 2005] Chantree, F., Kilgarriff, A., de Roeck, A., and Willis, A. (2005). Disambiguating coordinations using word distribution information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

[Chater and Manning, 2006] Chater, N. and Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10:335–344.

[Chiarcos, 2011] Chiarcos, C. (2011). Evaluating salience metrics for context-adequate realization of discourse referents. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, pages 32–43, Nancy, France.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

[Clark and Wilkes-Gibbs, 1986a] Clark, H. and Wilkes-Gibbs, D. (1986a). Referring as a collaborative process. *Cognition*, 22:1–39.

[Clark and Marshall, 1981] Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A. K., Sag, I. A., and Webber, B. L., editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge, UK.

[Clark and Wilkes-Gibbs, 1986b] Clark, H. H. and Wilkes-Gibbs, D. (1986b). Referring as a collaborative process. *Cognition*, 22:1–39.

[Clarke et al., 2013] Clarke, A., Elsner, M., and Rohde, H. (2013). Where's wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4(329).

[Coco and Keller, 2009] Coco, M. I. and Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci)*, pages 274–279, Amsterdam.

[Cohen and Levesque, 1985] Cohen, P. R. and Levesque, H. J. (1985). Speech acts and rationality. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*.

[Constant, 2012] Constant, N. (2012). Witnessable quantifiers licence type-e meaning: evidence from contrastive topic, equatives and supplements. In *Proceedings of the 22st Conference on Semantics and Linguistic Theory (SALT-22)*, pages 286–306, Chicago.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley series in telecommunications, New York.

[Cristea et al., 1999] Cristea, D., Ide, N., Marcu, D., and Tablan, V. (1999). Discourse structure and coreference: an empirical study. In *Proceedings of the ACL-99 Workshop The Relation of Discourse/dialogue Structure and Reference*, pages 46–53.

[Croitoru et al., 2011] Croitoru, M., Guizol, L., and Leclère, M. (2011). On link validity in bibliographic knowledge bases. In *Advances in Computational Intelligence*, volume 297 (5), pages 380–389, Boston. Springer, Communications in Computer and Information Science.

[Croitoru and van Deemter, 2007] Croitoru, M. and van Deemter, K. (2007). A conceptual graph approach to the generation of referring expressions. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2456–2461, Hyderabad, India.

[Cumming, 2013] Cumming, S. (2013). Names. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, CSLI, spring 2013 edition.

[Daelemans and van den Bosch, 2005] Daelemans, W. and van den Bosch, A. (2005). *Memory-based Language Processing*. Cambridge University Press, Cambridge.

[Dale, 1988] Dale, R. (1988). *Generating Referring Expressions in a Domain of Objects and Processes*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

[Dale, 1989a] Dale, R. (1989a). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 68–75.

[Dale, 1989b] Dale, R. (1989b). Generating recipes: An overview of EPICURE. In *Proceedings of the 2nd European Workshop on Natural Language Generation*, Edinburgh, UK.

[Dale, 1992] Dale, R. (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. The MIT Press, Cambridge, Mass.

[Dale et al., 2005] Dale, R., Geldof, S., and Prost, J.-P. (2005). Using natural language generation in automatic route description. *Journal of Research and Practice in Information Technology*, 37(1):89–105.

[Dale and Haddock, 1991] Dale, R. and Haddock, N. (1991). Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association of Computational Linguists (EACL)*, pages 161–166, Berlin.

[Dale and Reiter, 1995] Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

[Davey, 1974] Davey, A. (1974). *The Formalisation of Discourse Production*. PhD thesis, University of Edinburgh, Edinburgh, Scotland.

[Davey, 1978] Davey, A. (1978). *Discourse Production*. Edinburgh University Press, Edinburgh.

[de Ruiter, 2000] de Ruiter, J. (2000). The production of gesture and speech. In McNeill, D., editor, *Language and Gesture*. Cambridge University Press.

[de Ruiter et al., 2012] de Ruiter, J., Bangerter, A., and Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.

[Dekker, 1998] Dekker, P. (1998). Speaker's reference, descriptions and information structure. *Journal of Semantics*, 15(3):305–334.

[Dell et al., 1997] Dell, G., Schwartz, M., Martin, N., Saffran, E., and Gagnon, D. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104:801–838.

[DeVault et al., 2004] DeVault, D., Rich, C., and Sidner, C. L. (2004). Natural language generation and discourse context: Computing distractor sets from the focus stack. In *Proceedings of the 17th International Meeting of the Florida Artificial Intelligence Research Society (FLAIRS)*, Miami Beach.

[DeVault and Stone, 2004] DeVault, D. and Stone, M. (2004). Interpreting vague utterances in context. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling) conference*, pages 236–241, Geneva.

[Devitt, 2004] Devitt, M. (2004). The case for referential descriptions. In Bezuidenhout and Reimer, editors, *Descriptions and Beyond*, pages 280–305. Oxford University Press, Oxford.

[Di Eugenio et al., 2000] Di Eugenio, B., Jordan, P. W., Thomason, R. H., and Moore, J. D. (2000). The agreement process: an empirical investigation of human-human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53:1017–1076.

[Donnellan, 1966] Donnellan, K. S. (1966). Reference and definite descriptions. *Philosophical Review*, 75:281–304. [Reprinted in J.F. Rosenberg and C. Travis, editors, *Readings in the Philosophy of Language*, 195–211, Prentice Hall, Englewood Cliffs, NJ, 1971].

[Elbourne, 2005] Elbourne, P. (2005). *Situations and Individuals*. MIT Press, Cambridge, Mass.

[Elmagarmid et al., 2007] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16.

[Engelhardt et al., 2006] Engelhardt, P., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54:554–573.

[Engelhardt et al., 2011] Engelhardt, P. E., Demiral, S. B., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An erp study. *Brain and Cognition*, 77:304–314.

[Engonopoulos et al., 2013] Engonopoulos, N., Villalba, M., Titov, I., and Koller, A. (2013). Predicting the resolution of referring expressions from user behavior. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Short Papers*, Seattle.

[Fabbrizio et al., 2008a] Fabbrizio, G. D., Stent, A. J., and Bangalore, S. (2008a). Referring expression generation using speaker-based attribute selection and trainable realization (ATTREG). In *Proceedings of the 5th International Conference on Natural Language Generation (INLG'08)*, pages 211–214.

[Fabbrizio et al., 2008b] Fabbrizio, G. D., Stent, A. J., and Bangalore, S. (2008b). Trainable speaker-based referring expression generation. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL'08)*, pages 151–158.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning About Knowledge*. MIT Press, Cambridge, Mass.

[Fang et al., 2008] Fang, F., Boyaci, H., Kersten, D., and Murray, S. O. (2008). Attention-dependent representation of a size illusion in human v1. *Current Biology*, 18(21):1707–1712.

[Fang et al., 2014] Fang, R., Doering, M., and Chai, J. Y. (2014). Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI14)*, Quebec City.

[Fano, 1961] Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, Mass.

[Feldman, 1980] Feldman, A. M. (1980). *Welfare Economics and Social Choice Theory*. Kluwer, Boston.

[Fernández, 2009] Fernández, R. (2009). Salience and feature variability in definite descriptions with positive-form vague adjectives. In *Proceedings Workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*, Amsterdam.

[FitzGerald et al., 2013] FitzGerald, N., Artzi, Y., and Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1914–1925, Seattle, Washington.

[Ford and Olson, 1975] Ford, W. and Olson, D. (1975). The elaboration of the noun phrase in children's object descriptions. *Journal of Experimental Child Psychology*, 19:371–382.

[Frank and Goodman, 2012] Frank, M. and Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, 336:998.

[Frege, 1960] Frege, G. (1892 (1960)). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of Gottlob Frege*. Basil Blackwell, Oxford.

[Funakoshi et al., 2004] Funakoshi, K., Watanabe, S., Kuriyama, N., and Takunaga, T. (2004). Generating referring expressions using perceptual groups. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG)*, pages 51–61, Brockenhurst, UK.

[Fussell and Krauss, 1989] Fussell, S. R. and Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25(3):203–219.

[Fussell and Krauss, 1992] Fussell, S. R. and Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, 62:378–391.

[Gardent, 2002] Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 96–103, Philadelphia.

[Gardent and Striegnitz, 2007] Gardent, C. and Striegnitz, K. (2007). Generating bridging definite descriptions. In Bunt, H. and Muskens, R., editors, *Computing Meaning, Volume 3*, pages 369–396. Studies in Linguistics and Philosophy, Springer Publishers.

[Garoufi and Koller, 2014] Garoufi, K. and Koller, A. (2014). Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience*, 29(8):986–1001.

[Garrett, 1984] Garrett, M. (1984). The organization of processing structure for language production: Application to aphasic speech. In Caplan, D., Lecours, A., and Smith, A., editors, *Biological Perspectives on Language*, pages 172–193. MIT Press.

[Gatt, 2007] Gatt, A. (2007). *Generating coherent references to multiple entities*. PhD thesis, Department of Computing Science, University of Aberdeen.

[Gatt and Belz, 2010] Gatt, A. and Belz, A. (2010). Introducing shared task evaluation to nlg: The TUNA shared task evaluation challenges. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, pages 264–293. Springer Verlag, Berlin.

[Gatt et al., 2012] Gatt, A., Gompel, R., E.Krahmer, and Deemter, K. (2012). Does domain size impact speech onset time during reference production? In *Proceedings of the 34st Annual Conference of the Cognitive Science Society (CogSci)*, pages 1584–1589, Sapporo.

[Gatt et al., 2013a] Gatt, A., Krahmer, E., van Gompel, R., and van Deemter, K. (2013a). Production of referring expressions: Preference trumps discrimination. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 483–488, Berlin.

[Gatt and van Deemter, 2007] Gatt, A. and van Deemter, K. (2007). Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, 16:423–443.

[Gatt et al., 2007] Gatt, A., van der Sluis, I., and van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG)*, Schloss Dagstuhl, Germany.

[Gatt et al., 2011] Gatt, A., van Gompel, R., Krahmer, E., and van Deemter, K. (2011). Nondeterministic attribute selection in reference production. In *Proceedings of the CogSci workshop Production of Referring Expressions*, Boston.

[Gatt et al., 2013b] Gatt, A., van Gompel, R., van Deemter, K., and Krahmer, E. (2013b). Are we bayesian referring expression generators? In *Proceedings of the Cogsci workshop on Production of Referring Expressions. Associated with the 35th Annual Conference of the Cognitive Science Society*, Berlin.

[Geng and Hamilton, 2006] Geng, L. and Hamilton, H. (2006). Interestingness measures for data mining: a survey. *ACM Computing Surveys*, 38(3):1–32.

[Gibbon et al., 2000] Gibbon, D., Mertins, I., and Moore, R. K. (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology, and Product Evaluation*. Kluwer, Norwell (Mass) and Dordrecht, The Netherlands.

[Gibbs and Orden, 2012] Gibbs, R. and Orden, G. V. (2012). Pragmatic choice in conversation. *Topics in Cognitive Science*, 4(1).

[Glucksberg et al., 1966] Glucksberg, S., Krauss, R. M., and Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, 3(4):333–342.

[Gorniak and Roy, 2004] Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

[Goudbeek and Krahmer, 2012] Goudbeek, M. and Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, 4(2):166–183.

[Green and van Deemter, 2011] Green, M. J. and van Deemter, K. (2011). Vagueness as cost reduction: an empirical test. In *Proceedings of the Workshop on Production of Referring Expressions at the 33rd Annual Meeting of the Cognitive Science Society*, Boston.

[Grice, 1969] Grice, P. (1969). Utterer's meaning and intentions. *The Philosophical Review*, 68:147–177.

[Grice, 1975] Grice, P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 43–58. Academic Press, New York.

[Grishman and Sundheim, 1995] Grishman, R. and Sundheim, B. (1995). Design of the muc-6 evaluation. In *MUC-6*, pages 1–11, San Francisco.

[Groenendijk et al., 1996] Groenendijk, J., Stokhof, M., and Veltman, F. (1996). Coreference and modality. In *The Handbook of Contemporary Semantic Theory*, pages 179–214. Blackwell, Cambridge, Mass.

[Guhe, 2012] Guhe, M. (2012). Utility-based generation of referring expressions. *Topics in Cognitive Science*, 4(2):306–329.

[Guhe and Bard, 2012] Guhe, M. and Bard, E. (2012). Adapting referring expressions to the task environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 2404–2409, Austin, TX.

[Guhe and Bard, 2008] Guhe, M. and Bard, E. G. (2008). Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2404–2409, Austin, TX.

[Gundel et al., 1993] Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

[Gupta and Stent, 2005] Gupta, S. and Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Copora in Natural Language Generation (UCNLG)*, pages 1–6, Brighton, UK.

[Hajičová et al., 1998] Hajičová, E., Partee, B., and Sgall, P. (1998). *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Studies in Linguistics and Philosophy, Dordrecht.

[Hawkins, 1978] Hawkins, J. (1978). *Definiteness and Indefiniteness: A Study of Reference and Grammaticality Prediction*. Croom Helm, London.

[Hawthorne and Manley, 2012] Hawthorne, J. and Manley, D. (2012). *The Reference Book*. Oxford University Press, Oxford, UK.

[Heeman and Hirst, 1995] Heeman, P. A. and Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

[Heller et al., 2012] Heller, D., Skovbroten, K., and Tanenhaus, M. (2012). To name or to describe: shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2):166–183.

[Henschel et al., 2000] Henschel, R., Cheng, H., and Poesio, M. (2000). Pronominalisation revisited. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 306–312, Saarbrücken, Germany.

[Hermann and Deutsch, 1976] Hermann, T. and Deutsch, W. (1976). *Psychologie der Objektbenennung*. Huber Verlag, Bern.

[Hintzmann, 1990] Hintzmann, D. (1990). Human learning nd memory: connections and dissociations. *Annual Review of Psychology*, 41(1):109–139.

[Hirschmann and Chinchor, 1997] Hirschmann, L. and Chinchor, N. (1997). MUC-7 coreference task definition. In *MUC-7 Proceedings, Science Applications International Corporation*, Nancy, France.

[Hodges et al., 1996] Hodges, J., Yie, S., Reighart, R., and Boggess, L. (1996). An automated system that assists in the generation of document indexes. *Natural Language Engineering*, 2(2):137–160.

[Hofstadter and Sander, 2013] Hofstadter, D. and Sander, E. (2013). *Analogy as the Fuel and Fire of Thinking*. Basic Books, New York.

[Holden and van Orden, 2009] Holden, J. and van Orden, G. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological Review*, 2:318–342.

[Hopcroft, 1971] Hopcroft, J. (1971). An n log(n) algorithm for minimizing states in a finite automaton. In Kohave, Z., editor, *Theory of Machines and computations*. Academic Press.

[Horacek, 1996] Horacek, H. (1996). A new algorithm for generating referring expressions. In *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI)*, pages 577–581, Budapest, Hungary.

[Horacek, 1997] Horacek, H. (1997). An algorithm for generating referential descriptions with flexible interfaces. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 206–213, Madrid.

[Horacek, 2004] Horacek, H. (2004). On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 70–79, Brockenhurst, UK.

[Horrocks et al., 2006] Horrocks, I., Kutz, O., and Sattler, U. (2006). The even more irresistible SROIQ. In *Proc. of 10th Int. Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, Lake District of the UK.

[Horton and Keysar, 1996] Horton, W. S. and Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59:91–117.

[Howson and Urbach, 1996] Howson, C. and Urbach, P. (1996). *Scientific Reasoning: the Bayesian Approach (2nd edition)*. Open Court Publishing Company, Chicago and La Salle, Illinois.

[Huang and Snedeker, 2009] Huang, Y. T. and Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantic-pragmatics interface. *Cognitive Psychology*, 58:376–415.

[Jameson, 1983] Jameson, A. (1983). Impression monitoring in evaluation-oriented dialog: The role of the listener's assumed expectations and values in the generation of informative statements. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 616–620, Karlsruhe.

[Johnson-Laird, 2006] Johnson-Laird, P. (2006). *How We Reason*. Oxford University Press, Oxford.

[Jordan, 2000a] Jordan, P. W. (2000a). Can nominal expressions achieve multiple goals? an empirical study. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 142–149, Hong Kong.

[Jordan, 2000b] Jordan, P. W. (2000b). Influences on attribute selection in redescriptions: A corpus study. In *Proceedings of the Cognitive Science Conference*.

[Jordan, 2000c] Jordan, P. W. (2000c). *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. PhD thesis, University of Pittsburgh.

[Jordan, 2002] Jordan, P. W. (2002). Contextual influences on attribute selection for repeated descriptions. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Reference and Presupposition in Natural Language Generation and Understanding*. CSLI Publications, Stanford, Calif.

[Jordan and Walker, 2005] Jordan, P. W. and Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

[Jucks et al., 2008] Jucks, R., Becker, B.-M., and Bromme, R. (2008). Lexical entrainment in written discourse: Is experts' word use adapted to the addressee? *Discourse Processes*, 45(6):497–518.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (second edition)*. Pearson, Upper Saddle River, NJ.

[Kabasenche et al., 2012] Kabasenche, W., O'Rourke, M., and (Eds.), M. S. (2012). *Reference and Referring*. MIT Press, Cambridge, Mass.

[Kahneman, 2012] Kahneman, D. (2012). *Thinking Fast and Slow*. Penguin Books, London.

[Kamp and Reyle, 1993] Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic, and Discourse Representation Theory*. Kluwer, Studies in Linguistics and Philosophy 42, Dordrecht.

[Karlsson, 2007] Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.

[Kazemzadeh et al., 2014] Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). Referit game: Referring to objects in photographs of natural scenes. In *Proceedings of Conference on Empirical Methods in Natural language Processing (EMNLP2014*, Doha, Qatar.

[Kelleher and Kruijff, 2006a] Kelleher, J. and Kruijff, G.-J. (2006a). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1041–1048, Sydney, Australia.

[Kelleher and Kruijff, 2006b] Kelleher, J. D. and Kruijff, G.-J. (2006b). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL/COLING-06*.

[Kennedy, 1999] Kennedy, C. (1999). *Projecting the Adjective: the Syntax and Semantics of Gradability and Comparison*. PhD. Thesis, University of California Press.

[Keysar et al., 2000] Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

[Keysar et al., 2003] Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89:25–41.

[Khan, 2013] Khan, I. H. (2013). *Managing Ambiguity: Studies in Generation of Referring Expressions*. Lambert Academic Publishing, New York.

[Khan, 2015] Khan, I. H. (2015). Production of referring expressions in arabic. *International Journal of Speech Technology*.

[Khan et al., 2006] Khan, I. H., Ritchie, G., and van Deemter, K. (2006). The clarity–brevity trade–off in generating referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG)*, pages 89–91, Sydney, Australia.

[Khan et al., 2012] Khan, I. H., van Deemter, K., and Ritchie, G. (2012). Managing ambiguity in reference generation: the role of surface structure. *Topics in Cognitive Science*, 4(2):211–231.

[Kibble and Power, 2004] Kibble, R. and Power, R. (2004). Optimizing referential coherence in text generation. *Computational Linguistics*, 30:401–416.

[Kilgarriff, 2003] Kilgarriff, A. (2003). Thesauruses for natural language processing. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLPK)*, pages 5–13.

[Koller et al., 2010a] Koller, A., Gargett, A., and Garoufi, K. (2010a). A scalable model of planning perlocutionary acts. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (PozDial)*, Poznan.

[Koller and Stone, 2007] Koller, A. and Stone, M. (2007). Sentence generation as a planning problem. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Conference Proceedings (ACL)*, pages 337–343, Prague.

[Koller et al., 2010b] Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2010b). The first challenge on generating instructions in virtual environments. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *LNAI*. Springer.

[Koller et al., 2011] Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2011). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG)*, Nancy, France.

[Koolen et al., 2009] Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2009). Need I say more? On factors causing referential overspecification. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging Computational and Psycholinguistic Approaches (*PRE-COGSCI'09*)*.

[Koolen et al., 2013a] Koolen, R., Goudbeek, M., and Krahmer, E. (2013a). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37(2):395–411.

[Koolen et al., 2013b] Koolen, R., Krahmer, E., and Swerts, M. (2013b). The impact of bottom-up and top-down saliency cues on reference production. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci)*, pages 817–822, Berlin.

[Krahmer et al., 2012] Krahmer, E., Koolen, R., and Theune, M. (2012). Is it that difficult to find a good preference order for the incremental algorithm? *Cognitive Science*, 36(5):837–841.

[Krahmer and Theune, 1998] Krahmer, E. and Theune, M. (1998). Context sensitive generation of descriptions. In *ICSLP–98*, pages 1151–1154, Sydney, Australia.

[Krahmer and Theune, 2002] Krahmer, E. and Theune, M. (2002). Efficient context–sensitive generation of descriptions in context. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Givenness and Newness in Language Processing*, pages 223–264, CSLI Publications, CSLI, Stanford.

[Krahmer and Theune, 2010] Krahmer, E. and Theune, M. (2010). *Empirical Methods in Natural Language Generation*. Springer, Berlin and Heidelberg and New York.

[Krahmer et al., 2008] Krahmer, E., Theune, M., Viethen, J., and Hendrickx, I. (2008). Graph: the costs of redundancy in referring expressions. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 227–229, Salt Fork, Ohio.

[Krahmer and Van Deemter, 2012] Krahmer, E. and Van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Computational Linguistics*, 38(1):173–218.

[Krahmer et al., 2003] Krahmer, E., van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

[Kripke, 1977] Kripke, S. (1977). Speaker's reference and semantic reference. In French, P., Uehling, T., and Wettstein, H. K., editors, *Contemporary Perspectives in the Philosophy of Language*, pages 6–27. University of Minnesota Press, Minneapolis.

[Kripke, 1980] Kripke, S. (1980). *Naming and Necessity*. Harvard University Press, Cambridge, Mass.

[Kroch, 2000] Kroch, A. (2000). Syntactic change. In Baltin, M. and Collins, C., editors, *Handbook of Contemporary Syntactic Theory*. Blackwell, Oxford.

[Kronfeld, 1989] Kronfeld, A. (1989). Conversationally relevant descriptions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

[Kronfeld, 1990] Kronfeld, A. (1990). *Reference and Computation: An Essay in Applied Philosophy of Language*. Cambridge University Press, Cambridge.

[Kumar, 1992] Kumar, V. (1992). Algorithms for constraint satisfaction problems: a survey. *Artificial Intelligence Magazine*, 1:32–44.

[Kutlak, 2014] Kutlak, R. (2014). *Generation of Referring Expressions for an Unknown Audience*. PhD thesis, University of Aberdeen.

[Kutlak et al., 2011] Kutlak, R., van Deemter, K., and Mellish, C. (2011). Audience design in the generation of references to famous people. In *Proceedings of the 33th Meeting of the Cognitive Science Society*.

[Kutlak et al., 2012] Kutlak, R., van Deemter, K., and Mellish, C. (2012). Corpus-based metrics for assessing communal common ground. In *Proceedings of the 34th Meeting of the Cognitive Science Society*.

[Kutlak et al., 2013] Kutlak, R., van Deemter, K., and Mellish, C. (2013). Generation of referring expressions in large domains. In *Proceedings of the workshop Production of Referring Expressions, associated with the 35th Meeting of the Cognitive Science Society*.

[Kyburg and Morreau, 2000] Kyburg, A. and Morreau, M. (2000). Fitting words: Vague language in context. *Linguistics and Philosophy*, 23:577–579.

[Laird, 2012] Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press, Cambridge, Mass.

[Landau and Jackendoff, 1993] Landau, B. and Jackendoff, R. (1993). what17and where17in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265.

[Lane et al., 2006] Lane, L. W., Groisman, M., and Ferreira, V. S. (2006). Don't talk about pink elephants! : Speakers' control over leaking private information during language production. *Psychological Science*, 17:273–277.

[Langkilde and Knight, 1998] Langkilde, I. and Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th COLING and 36th ACL Conference*, pages 704–710, Montreal, Quebec.

[Larson and Segal, 1995] Larson, R. and Segal, G. (1995). *Knowledge and Meaning. An Introduction to Semantic Theory.* MIT Press, Cambridge, Mass.

[Levelt, 1989] Levelt, W. (1989). *Speaking*. The MIT Press, Cambridge, Mass.

[Levelt et al., 1999] Levelt, W., Roelofs, A., and Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75.

[Levinson, 1983] Levinson, S. (1983). *Pragmatics*. Cambridge University Press, Cambridge.

[Lewandowsky and Farrell, 2011] Lewandowsky, S. and Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Open Court Publishing Company, Los Angeles.

[Lewis, 1969] Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press, Cambridge, Mass.

[Lewis, 1979] Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359.

[Li et al., 2011] Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, pages 220–228, Portland, Oregon.

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL98)*, pages 768–774, Montreal.

[Lin, 1998b] Lin, D. (1998b). An information–theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 296–304, Madison, Wisconsin.

[Linsky, 1971] Linsky, L. (1971). *Reference and Modality*. Oxford University Press, Oxford.

[Lipman, 2009] Lipman, B. (2009). Why is language vague? In *Working Papers*, Department of Economics, Boston University.

[Lønning, 1997] Lønning, J. T. (1997). Plurals and collectivity. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 1009–1054. Elsevier, Amsterdam.

[Ludlow and Neale, 1991] Ludlow, P. and Neale, S. (1991). Indefinite descriptions: in defense of russell. *Linguistics and Philosophy*, 14:171–202.

[Ludlow and Segal, 2004] Ludlow, P. and Segal, G. (2004). On a unitary semantical analysis for definite and indefinite descriptions. In Reimer and Bezuidenhout, editors, *Descriptions and Beyond*, pages 420–436. Oxford University Press, Oxford, UK.

[Malouf, 2000] Malouf, R. (2000). The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–92.

[Matthews et al., 2007] Matthews, D. E., Lieven, E. V. M., and Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, 78:1744–1759.

[Matuszek et al., 2006] Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering*.

[McCarthy, 1980] McCarthy, J. (1980). Circumscription: a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39.

[McCluskey, 1965] McCluskey, E. J. (1965). *Introduction to the Theory of Switching Circuits*. McGraw-Hill, New York.

[McCoy and Strube, 1999] McCoy, K. and Strube, M. (1999). Generating anaphoric expressions: pronoun or definite description? In *Proceedings of ACL Workshop on Discourse and Reference Structure*, pages 63–71, University of Maryland, College Park.

[Mellish et al., 2006] Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., and Reape, M. (2006). A reference architecture for natural language generation systems. *Natural Language Engineering*, 12:1–34.

[Meteer, 1991] Meteer, M. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7:296–304.

[Metzing and Brennan, 2003] Metzing, C. A. and Brennan, S. E. (2003). When conceptual pacts are broken: partner effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:201–213.

[Mill, 1843] Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. John W. Parker, London.

[Milsark, 1977] Milsark, G. (1977). Towards an explanation of certain peculiarities of the existential construction in english. *Linguistic Analysis*, 3(1):1–30.

[Mitchell, 2012] Mitchell, M. (2012). *Generating Reference to Visual Objects*. PhD. Thesis, University of Aberdeen, Aberdeen.

[Mitchell et al., 2011a] Mitchell, M., Dunlop, A., and Roark, B. (2011a). Semi-supervised modeling for prenominal modifier ordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 236–241, Portland.

[Mitchell et al., 2012] Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., and Daume, H. (2012). Midge: generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

[Mitchell et al., 2013a] Mitchell, M., Reiter, E., and van Deemter, K. (2013a). Attributes in visual object reference. In *Proceedings of the Cogsci workshop on Production of Referring Expressions. Associated with the 35th Annual Conference of the Cognitive Science Society*.

[Mitchell et al., 2013b] Mitchell, M., Reiter, E., and van Deemter, K. (2013b). Typicality and object reference. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

[Mitchell et al., 2010] Mitchell, M., van Deemter, K., and Reiter, E. (2010). Natural reference to objects in a visual domain. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*.

[Mitchell et al., 2011b] Mitchell, M., van Deemter, K., and Reiter, E. (2011b). Applying machine learning to the choice of size modifiers. In *Proceedings of the Cogsci workshop on Production of Referring Expressions. Associated with the 33th Annual Conference of the Cognitive Science Society*.

[Mitchell et al., 2011c] Mitchell, M., van Deemter, K., and Reiter, E. (2011c). Two approaches for generating size modifiers. In *Proceedings of the 13th European Workshop on Natural Language Generation*.

[Mitchell et al., 2013c] Mitchell, M., van Deemter, K., and Reiter, E. (2013c). Generating expressions that refer to visible objects. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[Mitkov, 2002] Mitkov, R. (2002). *Anaphora Resolution*. Longman, London.

[Mostowski, 1957] Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae*, 44:235–273.

[Motwani and Raghavan, 1995] Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press, Cambridge, UK.

[Murray et al., 2006] Murray, S. O., Boyaci, H., and Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience*, 9(3):429–434.

[Nadig et al., 2015] Nadig, A., Seth, S., and Sasson, M. (2015). Global similarities and multifaceted differences in the production of partner-specific referential pacts by adults with autism spectrum disorders. *Frontiers in Psychology*, 6(1888):1–14.

[Nash, 1950] Nash, J. (1950). The bargaining problem. *Econometrica*, 18:155–162.

[Neale, 1990] Neale, S. (1990). *Descriptions*. MIT Press, Cambridge (Mass.).

[Newcombe et al., 1959] Newcombe, H. B., Kennedy, J. M., S. J. Axford, S., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 3130:954–959.

[Nickerson et al., 1987] Nickerson, R. S., Baddeley, A., and Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, 64(3):245–259.

[Nieuwland and van Berkum, 2008] Nieuwland, M. and van Berkum, J. (2008). The neurocognition of referential ambiguity in language comprehension. *Language and Linguistics Compass*, 2(4):603–630.

[Nozari et al., 2014] Nozari, N., Arnold, J., and Thompson-Schill, S. (2014). The effects of anodal stimulation of the left prefrontal cortex on sentence production. *Brain Stimulation*, 7(6):784–792.

[Nunberg, 1978] Nunberg, G. (1978). *The Pragmatics of Reference*. PhD Thesis, City University of New York, New York.

[Oberlander, 1998] Oberlander, J. (1998). Do the right thing ... but expect the unexpected. *Computational Linguistics*, 24(3):501–507.

[O'Donnell et al., 1998] O'Donnell, M., Cheng, H., and Hitzeman, J. (1998). Integrating referring and informing in np planning. In *Proceedings of the ACL Workshop on The Computational Treatment of Nominals*, pages 46–55, Montreal, Canada.

[Olson, 1970] Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77:257–273.

[Paraboni and van Deemter, 2014] Paraboni, I. and van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, 29(8):1002–1017.

[Paraboni et al., 2007] Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Computational Linguistics*, 33(2):229–254.

[Passonneau, 1995] Passonneau, R. (1995). Integrating Gricean and attentional constraints. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*.

[Passonneau, 1996] Passonneau, R. (1996). Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, 39:229–264.

[Pechmann, 1989] Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27:98–110.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet:Similarity – measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, pages 1024–1025.

[Peters and Westerstahl, 2006] Peters, S. and Westerstahl, D. (2006). *Quantifiers in Language and Logic*. Oxford University Press, Oxford, UK.

[Pickering and Garrod, 2004] Pickering, M. and Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27:169–226.

[Pierrehumbert and Hirschberg, 1990] Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions and Plans in Communication and Discourse*, pages 271–311. MIT Press, Cambridge, Mass.

[Pinkal, 1979] Pinkal, M. (1979). How to refer with vague descriptions. In Bäuerle, A., Egli, U., and von Stechow, A., editors, *Semantics from Different Points of View*, pages 32–50. Springer, Berlin.

[Piwek, 2008] Piwek, P. (2008). Proximal and distal in language and cognition: Evidence from deictic demonstratives in Dutch. *Journal of Pragmatics*, 40:694–718.

[Piwek, 2009] Piwek, P. (2009). Salience and pointing in multimodal reference. In *Proceedings of Production of Referring Expressions: bridging the gap between computational and empirical approaches to generating reference (PRE-CogSci'09)*.

[Poesio et al., 2004] Poesio, M., Stevenson, R., di Eugenio, B., and Hitzeman, J. (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30:309–363.

[Poesio and Vieira, 1998] Poesio, M. and Vieira, R. (1998). A corpus–based investigation of definite description use. *Computational Linguistics*, 24:183–216.

[Pouget et al., 2013] Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178.

[Putnam, 1975] Putnam, H. (1975). The meaning of 'meaning'. In Gunderson, K., editor, *Language, Mind, and Knowledge*, pages 131–193. University of Minnesota Press, Minnesota Studies in the Philosophy of Science (Volume 7), Minneapolis.

[Raghunathan, 2013] Raghunathan, B. (2013). *The Complete Book of Data Anonymization: From Planning to Implementation*. CRC Press, Boca Raton, London, and New York.

[Recanati, 1993] Recanati, F. (1993). *Direct Reference: From Language to Thought*. Blackwell, Oxford.

[Reimer and Bezuidenhout, 2004] Reimer, M. and Bezuidenhout, A. (2004). *Descriptions and Beyond*. Oxford University Press, Oxford.

[Reiter, 1990a] Reiter, E. (1990a). The computational complexity of avoiding conversational implicatures. In *Proc. 28th Annual Meeting of the Association for Computational Linguistics.*

[Reiter, 1990b] Reiter, E. (1990b). The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 97–104.

[Reiter, 2007] Reiter, E. (2007). The shrinking horizons of computational linguistics. *Computational Linguistics*, 21:283–290.

[Reiter and Dale, 2000] Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

[Ren et al., 2010] Ren, Y., van Deemter, K., and Pan, J. (2010). Charting the potential of Description Logic for the generation of referring expressions. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 115–124.

[Ritchie, 1986] Ritchie, G. (1986). A rational reconstruction of the Proteus sentence planner. In *Proceedings of 10th International Conference on Computational Linguistics/22nd Annual Meeting of the Association for Computational Linguistics*, pages 327–329.

[Robinson and Voronkov, 2001] Robinson, A. and Voronkov, A. (2001). *Handbook of Automated Reasoning Volumes 1 and 2*. Elsevier and MIT Press.

[Rosch, 1978] Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. L., editors, *Cognition and Categorization*, pages 27–48. Erlbaum, Hillsdale, NJ.

[Russell, 1905] Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

[Russell and Norvig, 2003] Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach, 2nd edition*. Prentice–Hall, Englewood Cliffs, NJ.

[Ryu et al., 2010] Ryu, J., Jung, Y., Kim, K., and Myaeng, S. (2010). Automatic extraction of human activity knowledge from method-describing web articles. In *Proceedings of the 1st Workshop on Automated Knowledge Base Construction*, pages 16–23.

[Scha and Stallard, 1988] Scha, R. and Stallard, D. (1988). Multi-level plurals and distributivity. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 17–24, Buffalo, NY.

[Schmuckler, 2001] Schmuckler, M. A. (2001). What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

[Schwartzschild, 2002] Schwartzschild, R. (2002). Singleton indefinites. *Journal of Semantics*, 19(3):289–314.

[Schwarzkopf et al., 2010] Schwarzkopf, D. S., Song, C., and Rees, G. (2010). The surface area of human v1 predicts the subjective experience of object size. *Nature Neuroscience*, 14(1):28–30.

[Searle, 1969] Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.

[Sedivy, 2003] Sedivy, J. (2003). Pragmatic versus form–based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1):3–23.

[Sedivy, 2007] Sedivy, J. C. (2007). Implicature during real time conversation: a view from language processing research. *Philosophy compass*, 2/3:275–496.

[Sedivy et al., 1999] Sedivy, J. G., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71:109–147.

[Semenza and Zettin, 1989] Semenza, C. and Zettin, M. (1989). Evidence from aphasia for the role of proper names as pure referring expressions. *Nature*, 342:678–679.

[Seylan et al., 2009] Seylan, I., Franconi, E., and de Bruijn, J. (2009). Effective query rewriting with ontologies over dboxes. In *International Joint Conference on Artificial Intelligence (IJCAI2009)*, pages 923–930, Pasadena, Calif.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

[Shaw and McKeown, 2000] Shaw, J. and McKeown, K. R. (2000). Generating referring quantified expressions. In *Proceedings of the 1st International Conference on Natural Language Generation*.

[Siddharthan and Copestake, 2004] Siddharthan, A. and Copestake, A. (2004). Generating referring expressions in open domains. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 407–414, Barcelona, Spain.

[Siddharthan et al., 2011] Siddharthan, A., Nenkova, A., and McKeown, K. (2011). Information status distinctions and referring expressions: an empirical study of references to people in news summaries. *Computational Linguistics*, 37(4).

[Stalnaker, 1973] Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, 2:447–457.

[Stalnaker, 1978] Stalnaker, R. (1978). Pragmatic presuppositions. In Munitz, M. and Unger, P., editors, *Semantics and Philosophy*, pages 197–213. New York University Press, New York.

[Stoia et al., 2006] Stoia, L., Shockley, D. M., Byron, D. K., and Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG)*, pages 81–88, Sydney, Australia.

[Stone, 2000] Stone, M. (2000). On identifying sets. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG)*, pages 116–123, Mitzpe Ramon.

[Stone et al., 2003] Stone, M., Doran, C., Webber, B., Bleam, T., and Palmer, M. (2003). Microplanning with communicative intentions: The SPUD system. *Computational Intelligence*, 19(4):311–381.

[Stone and Webber, 1998] Stone, M. and Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In *Proceedings of the 9th International Workshop on Natural Language Generation (INLG)*, pages 178–187, Niagara–on–the–Lake, Ontario.

[Strawson, 1959] Strawson, P. (1959). *Individuals: an Essay in Descriptive Metaphysics*. Methuen, London.

[Strawson, 1950] Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.

[Sun, 2008] Sun, R. (2008). *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, Cambridge.

[Taleb, 2010] Taleb, N. N. (2010). *The Black Swan: the impact of the highly improbable (2nd ed.)*. Penguin, London.

[Theune et al., 2011] Theune, M., Koolen, R., Krahmer, E., and Wubben, S. (2011). Does size matter: how much data is required to train a reg algorithm? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL2011)*, pages 660–664, Portland, Oregon.

[Theune et al., 2007] Theune, M., Touset, P., Viethen, J., and Krahmer, E. (2007). Cost-based attribute selection for generating referring expressions (GRAPH-FP and GRAPH-SC). In *Proceedings of UCNLG+MT: Language Generation and Machine Translation*, pages 95–97.

[Thomas and McClelland, 2008] Thomas, M. and McClelland, J. (2008). Connectionist models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press.

[Thórisson, 1994] Thórisson, K. R. (1994). Simulated perceptual grouping: an application to human-computer interaction. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, pages 876–881, Atlanta, Georgia.

[Tintarev and Masthoff, 2012] Tintarev, N. and Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems: methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction*, 22(4-5):399–439.

[Treisman and Gelade, 1980] Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

[Turing, 1950] Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX(2236):433–460.

[Turner et al., 2009] Turner, R., Sripada, S., and Reiter, E. (2009). Generating approximate geographic descriptions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 42–49, Athens, Greece.

[Turner et al., 2008] Turner, R., Sripada, S., Reiter, E., and Davy, I. P. (2008). Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 16–24.

[van Benthem, 1986] van Benthem, J. (1986). *Essays In Logical Semantics*. Reidel, Dordrecht.

[van Deemter, 2000] van Deemter, K. (2000). Generating vague descriptions. In *Proceedings of the 1st International Natural Language Generation Conference (INLG)*, pages 179–185, Mitzpe Ramon.

[van Deemter, 2002] van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the Incremental Algorithm. *Computational Linguistics*, 28(1):37–52.

[van Deemter, 2004] van Deemter, K. (2004). Finetuning an nlg system through experiments with human subjects: The case of vague descriptions. In *Proceedings of the 3rd International Conference on Natural Language Generation, INLG-04*.

[van Deemter, 2006] van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

[van Deemter, 2009a] van Deemter, K. (2009a). Utility and language generation: the case of vagueness. *Journal of Philosophical Logic*, 38(6):607–632.

[van Deemter, 2009b] van Deemter, K. (2009b). What game theory can do for nlg: the case of vague language. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*.

[van Deemter, 2010] van Deemter, K. (2010). *Not Exactly: in Praise of Vagueness*. Oxford University Press, Oxford, UK.

[van Deemter, 2014] van Deemter, K. (2014). Referability. In Stent, A. and Bangalore, S., editors, *Natural Language Genetration in Interactive Systems*. Cambridge University Press.

[van Deemter and Gatt, 2007] van Deemter, K. and Gatt, A. (2007). Content determination in gre: evaluating the evaluator. In *Proceedings of UCNLG+MT: Language Generation and Machine Translation*.

[van Deemter et al., 2012a] van Deemter, K., Gatt, A., van der Sluis, I., and Power, R. (2012a). Assessing the incremental algorithm: a response to krahmer et al. *Cognitive Science*, 36(5):842–845.

[van Deemter et al., 2012b] van Deemter, K., Gatt, A., van der Sluis, I., and Power, R. (2012b). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5):799–836.

[van Deemter et al., 2012c] van Deemter, K., Gatt, A., van Gompel, R., and Krahmer, E. (2012c). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2).

[van Deemter and Halldórsson, 2001] van Deemter, K. and Halldórsson, M. M. (2001). Logical form equivalence: The case of referring expressions generation. In *Proceedings of the 8th European Workshop on Natural Language Generation (ENLG)*, Toulouse, France.

[van Deemter and Kibble, 2000] van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26:629–637.

[van Deemter and Kibble, 2002] van Deemter, K. and Kibble, R., editors (2002). *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI Publications, Stanford, Calif.

[van Deemter and Krahmer, 2007] van Deemter, K. and Krahmer, E. (2007). Graphs and Booleans: on the generation of referring expressions. In Bunt, H. and Muskens, R., editors, *Computing Meaning, Volume 3*, pages 397–422. Studies in Linguistics and Philosophy, Springer Publishers.

[van Deemter et al., 2006] van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (Special Session on Data Sharing and Evaluation), INLG-06*, pages 130–132, Sydney.

[van der Sluis et al., 2007] van der Sluis, I., Gatt, A., and van Deemter, K. (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 (RANLP)*.

[van der Sluis and Krahmer, 2004] van der Sluis, I. and Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the ICSLP-2004*, Jeju, Korea.

[van Gompel et al., 2012] van Gompel, R., Gatt, A., Krahmer, E., and van Deemter, K. (2012). Pro: A computational model of referential overspecification. In *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference*, Riva del Garda, Italy.

[van Gompel et al., 2014] van Gompel, R., Gatt, A., Krahmer, E., and van Deemter, K. (2014). Overspecification in reference: modelling size contrast effects. In *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference [abstract]*, Edinburgh.

[van Hentenryck, 1989] van Hentenryck, P. (1989). *Constraint Satisfaction in Logic Programming*. The MIT Press, Cambridge, Mass.

[van Langendonck, 2007] van Langendonck, W. (2007). *Theory and Typology of Proper Names*. Mouton de Gruyter, The Hague.

[van Nieuwland et al., 2007] van Nieuwland, M., Petersson, K. M., and Berkum, J. V. (2007). On sense and reference: examining the functional neuroanatomy of referential processing. *NeuroImage*, 37(3):993–1004.

[van Rij et al., 2013] van Rij, J., van Rijn, H., and Hendriks, P. (2013). How wm load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, 5(3):564–580.

[Vanderschraaf and Sillari, 2009] Vanderschraaf, P. and Sillari, G. (2009). Common knowledge. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, CSLI, spring 2009 edition.

[Vicente and Wang, 1998] Vicente, K. and Wang, J. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, 105(1):33–57.

[Viethen and Dale, 2006a] Viethen, J. and Dale, R. (2006a). Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation, INLG-06*.

[Viethen and Dale, 2006b] Viethen, J. and Dale, R. (2006b). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 63–70, Sydney, Australia.

[Viethen and Dale, 2008] Viethen, J. and Dale, R. (2008). The use of spatial relations in referring expressions. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, pages 59–67.

[Viethen and Dale, 2010] Viethen, J. and Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the 8th Australasian Language Technology Workshop*, pages 81–89, Melbourne.

[Viethen et al., 2008] Viethen, J., Dale, R., Krahmer, E., Theune, M., and Touset, P. (2008). Controlling redundancy in referring expressions. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

[Vigliocco and Hartsuiker, 2002] Vigliocco, G. and Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin*, 128(3):442–472.

[Wahlster and Kobsa, 1989] Wahlster, W. and Kobsa, A. (1989). User models in dialogue systems. In Kobsa, A. and Wahlster, W., editors, *User Models in Dialogue Systems*, pages 4–34. Springer Verlag, Berlin.

[Wang et al., 2013] Wang, Z., Busemeyer, J., Atmanspacher, H., and Pothos, E. (2013). The potential of using quantum theory to build models of cognition. *Topics in Cognitive Science*, 5(4):672–688.

[Waterfield, 1987] Waterfield, R. A. (1987). *Plato: Theaetetus*. Penguin Books, London.

[Westerbeek et al., 2015] Westerbeek, H., Koolen, R., and Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, 6 July 2015.

[Whitehurst, 1976] Whitehurst, G. J. (1976). The development of communication: changes with age and modeling. *Child Development*, 47(473–482).

[Wicklund, 2012] Wicklund, M. D. (2012). *Use of Referring Expressions by Autistic Children in Spontaneous Conversations: Does Impaired Metarepresentational Ability Affect Reference Production?* PhD thesis, University of Minnesota.

[Wilson, 1991] Wilson, D. (1991). Reference and relevance. In *UCL Working Papers in Linguistics*, pages 167–191, University College London, London.

[Wilson and Sperber, 2004] Wilson, D. and Sperber, D. (2004). Relevance theory. In Horn, L. and Ward, G., editors, *Blackwell's Handbook of Pragmatics*. Blackwell.

[Winograd, 1972] Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.

[Witten et al., 2011] Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, Vermont.

[Woodham-Smith, 1954] Woodham-Smith, C. (1954). *The Reason Why*. McGraw-Hill, London.

[Wu and Keysar, 2007] Wu, S. and Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, 31:169–181.

[Yang et al., 2011] Yang, Y., Teo, C. L., Daumé, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 444–454, Stroudsburg, Penn.

Index