Generating Coherent References to Multiple Entities

Albert Gatt

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of Aberdeen.

Department of Computing Science

2007

Declaration

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: 2007

Abstract

What constitutes an adequate reference to a set of objects? Despite intensive research on the Generation of Referring Expressions (GRE), many GRE algorithms either lack empirical backing, or are motivated by concerns which arguably shift their focus away from the crucial problem, which is to generate *natural* descriptions, much as a person would generate them in a comparable situation. This problem becomes much more pronounced in the case of plural reference, where even psycholinguistic research is lacking.

This thesis focuses on the generation of plurals, with particular attention to the semantic heart of the problem, that is, *content determination*. The empirical and computational work addresses two hypotheses. First, descriptions of sets or groups of entities are more adequate if they maximise the similarity between elements of a set. Second, the form and content of referring expressions are strongly determined by the way entities are categorised, that is, what ontological category they belong to.

The first three chapters set the stage with an in-depth theoretical and empirical evaluation of the state of the art in GRE. Here, three main contributions are made. The first is the construction of a *semantically transparent* corpus of singular and plural descriptions. Second, an empirical investigation into reference by human authors in this corpus sheds further light on the content determination problem. Third, an evaluation study is conducted on various existing algorithms. This study is unique in that it is the first to directly address the semantic issues while attempting to abstract away from linguistic realisation. Moreover, it focuses not only on singular, but also on plural descriptions.

The second part of the thesis focuses directly on plurals. It begins (Chapter 5) with a test of the similarity hypothesis on corpus data, leading to the development of a new algorithm which addresses the issues of similarity and conceptual categorisation. The algorithm is also extended with a form of aggregation, again motivated by a new corpus study. This work is generalised to pluralities in discourse in Chapter 6, which starts from the hypothesis that pluralities should be *conceptually coherent*, that is, should conceptualise entities from the same perspective. This hypothesis is investigated in a series of five psycholinguistic experiments. Finally, Chapter 7 uses the results of the previous two chapters to build an integrated framework for content determination in GRE. Among the contributions of this second part of the thesis are (a) the use of an experimental psycholinguistic methodology to test hypothesis that are relevant to generation; (b) the proposal of a novel approach to generation that seeks to satisfy conceptual coherence through the use of corpus-derived similarity metrics.

Acknowledgements

Over the three years that it took to complete this thesis, I was fortunate enough to have worked in two different research groups, the Information Technology Research Institute (ITRI) at the University of Brighton, and the NLG Group at the University of Aberdeen. Throughout this period, my principal supervisor, Kees van Deemter, was a constant reference point. Kees's combination of patience and humour, in conjunction with a capacity to ask the difficult questions without making them seem insurmountable, made him an ideal colleague, as well as a guide. It is only through discussions with him that what sometimes started out as a wild idea became a viable research direction. For that, and for much else, I am grateful.

My work at the ITRI during my first year benefited immensely from the input of Richard Power and Sebastian Varges, both of whom were involved in the TUNA project. Adam Kilgarriff's corpus-based lexicographic work also exerted a strong influence, which will be apparent in many of the following chapters. I also thank Paul Piwek, who acted as a willing sounding board for my ideas, and Donia Scott, who directed the ITRI and gave me the opportunity to work there.

I think that all the members of the NLG group at Aberdeen have contributed to this work, if only because of the friendly and intensively research-focused atmosphere they helped to generate. I am especially grateful to Ehud Reiter. As my second supervisor in Aberdeen, Ehud was often instrumental in shedding light on fruitful avenues of research. His uncanny knack for seeing the forest where others might see only trees has helped me shape many of my ideas. Discussions with Graeme Ritchie were invariably useful, while Ielka van der Sluis was an especially close collaborator on TUNA. I've also had many discussions with fellow research students. Thanks are especially due to Ross Turner and Imtiaz Hussain Khan, whose work turned out to have interesting intersections with mine.

Many other people contributed in some way. I am especially indebted to Emiel Krahmer for comments and insights on drafts of papers and chapters. Reviewers for a number of conferences and journals, which are listed in Appendix E, also made important contributions. I am grateful to Jon Oberlander and Jim Hunter, who reviewed my thesis proposal, and to Robert Dale and Chris Mellish for reading and examining the thesis and helping to make it a better piece of work.

Last, but certainly not least, I wish to thank my parents, Mary Grace and Alfred, for never trying to curb my curiosity. And Clare, for being there throughout and for helping me to stay sane. This contribution is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 A painting

You are standing in the Museo Prado in Madrid, before a painting by Velazquèz. *Las Meninas* (Figure 1.1) attracts the throngs and you have been unable to get close to the masterwork. Still, there's a reasonably good view, and then again, this is a painting one feels one knows well; you've seen it a thousand times in pictures. Your partner, who doesn't know much about art (he's a computer scientist), suddenly asks you the following question:

> Which of the figures is the *Infanta*?

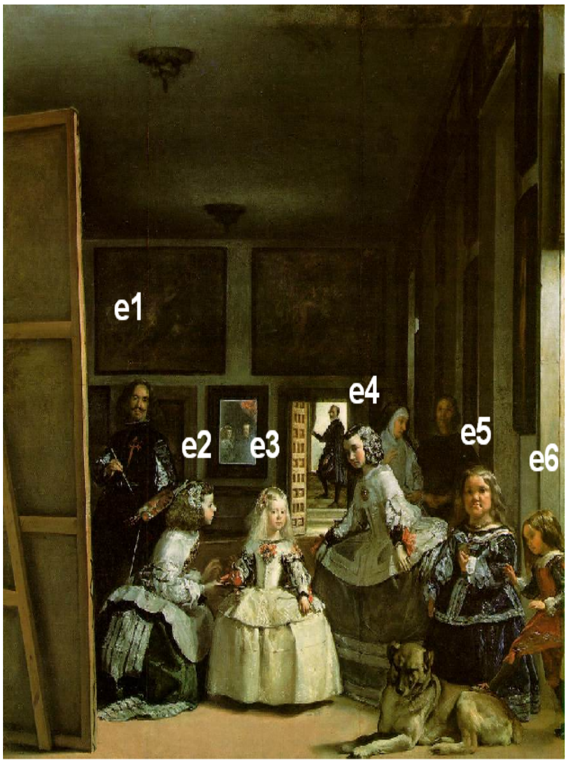


Figure 1.1: Las Meninas, by Velazquèz

The problem with questions like this is that they afford too many different answers, all of which could be equally 'satisfactory', in the sense that they would all single out the one figure in the painting which is (an image of) the young Princess of Spain. You could conceivably answer your partner's question by saying *the princess*. However, this would require that your partner know that there was one figure in the painting who was a princess, and that he be in a position to single it out. Similarly, a description such as *the painter's model* would beat the purpose: it would require your interlocutor's knowing which of the many figures in the painting was Velazquèz's model. Such descriptions might very well invite requests for clarification, so maybe something more detailed might serve to stave off the barrage in advance. *The leftmost girl but one* would probably do the trick, as would *the girl who is being tended by a maid*.

One of the potential pitfalls here is a mismatch between your knowledge state and your partner's. It's a safe bet that he can see whatever you can see, but you bring additional knowledge to bear which is not available perceptually (for instance, that some of the figures are maids of honour, that the painter in the painting is widely held to be Velazquèz himself, and so on). The perceptual input makes it more likely that some things will be more immediately available for use in your referring expression. Saying something like *the daughter of the King and Queen, seen in the mirror at the back* is playing with fire: Does your partner know that the two figures reflected in the mirror are in fact the King and Queen; indeed, can he even make them out clearly enough for this expression to be any use?

Although all of these **referring expressions** would satisfy the task at hand, namely, to identify the object or person your partner has asked about, this example points to a host of subsidiary motivations that could underlie your final choice. Some of these motivations might conflict. For instance, a throbbing hangover from the night before might force you to make your reply as curt as possible (forestalling further unwanted verbal interaction). Although *the princess* would do nicely, as would *the painter's model*, there is always the question of your partner's knowledge state. Furthermore, brevity comes at a cost, for it would require you to weigh your options carefully to find the shortest description that would rule out everyone in the painting, except for the girl in question. At the risk of giving out the wrong signal, verbosity might be cheaper.

There are other factors at play which you might not even be consciously aware of. Suppose you had mentioned, before coming to view the painting, that Velazquèz had indeed included himself, the King, and the Queen looking out of the painting, and that it featured a royal court with maids of honour. All this talk of royalty might bias you towards saying something like *the princess*, simply because the term is easiest to recall given that the topic is freshest in your mind. In short, you can talk about figures in the painting from several different points of view.

Presumably, things would get even worse if your partner's question required you to identify two or more objects. What would be a good reply to *Which of the figures are the maids?* Do you go about describing each of the figures, except for the painter and the princess (and, of course, the dog). Presumably, a description such as *the woman second from the left, the woman fourth from the left* … and so on is a possibility. So is negation: *Every woman in the painting except the middle one*. We've already seen that you can describe a single object from many points of view. When your intended referent is a plurality, it seems advisable to be consistent in the point of view you take on the objects. In answer to the question *Which of the figures are the maids?*, the following description, though possible, is not necessarily the best one:

> *The first woman from the left, the third woman from the left, the short woman in a black dress …*

Is it wise to begin by describing a woman from one point of view (her location), and talk about the height and dress of another woman in the same breath? The point of view you take on an object is itself informative, in that it suggests (to your partner in this case) that your choice of words is somehow relevant to your intention. Mixing perspectives might be puzzling for your interlocutor. It might also end up being more work for you, assuming that it involves relinquishing a particular perspective halfway through an utterance, and taking up a new one. But if consistency is a virtue, maintaining it may be a source of effort for you, since how you describe one object will depend on the way you're going to describe the others.

These considerations are beginning to seem rather daunting. And if, in addition to your overarching intention to identify the objects in question, context, continuity, and word choice are important factors, then you might even begin to consider the cheaper option of ignoring the question, at the expense of a row later on. It's a pity that, given the crowd surrounding the painting and your distance from it it's impossible to point.

1.2 The domain of inquiry

The process of unambiguously referring to an object is such a commonplace in everyday verbal behaviour, and seems to occur so effortlessly, that the above situation might seem over-dramatic. With the possible exception of throbbing hangovers, speakers seldom think long and hard about how to refer to objects. They probably only consider their intentions and/or choices in greater detail in case of referential failure, that is, when their interlocutors or target audience fail to identify the entities they have in mind.

A process that is ostensibly carried out effortlessly by humans is a good candidate for investigation by cognitive scientists. Given the complex interplay of factors in a referential situation, the apparent simplicity of the act itself suggests a good deal of underlying machinery of which we may only be aware to a limited extent. This makes it all the more challenging for a computer system that seeks to carry out the same process. Artificial Intelligence (AI) systems are often designed to emulate a specific human cognitive capacity. The aims of computational modelling may be practical. For example, an algorithm for the **Generation of Referring Expressions** (GRE) is often a requirement in **Natural Language Generation** (NLG), which is 'concerned with the construction of computer systems that can produce understandable text in English or other human languages from some underlying nonlinguistic representation of information' (Reiter and Dale, 1997, p.57). Since object description and identification is a pervasive aspect of linguistic communication, such algorithms contribute to the usefulness and communicative effectiveness of text.

Another aim of computational modelling of some human cognitive capacity is to obtain a better understanding of that capacity, as an object of study in its own right. As the example in the previous section suggested, a process such as referring may, under close scrutiny, give rise to questions about the underlying mechanisms, and further light could be shed on these issues by the design of formally explicit models to achieve the same aims, based on human performance in the same situations. This more 'cognitivist' bias is evident in another definition of NLG, as 'the process whereby thought is rendered into language' (McDonald, 2000, p.147). Though this definition seems to contain a view of NLG as a discipline akin to cognitive modelling (which is not the view taken here), it does suggest that there are interesting and non-trivial links between what humans do and what AI systems can be programmed to do.

These two possible objectives do not necessarily conflict. Indeed, it could be argued that they should be achieved in tandem, since a good way to model a human cognitive capacity, achieving a practical and viable technological solution, is to study the way it is performed by humans. As one author put it, ‘language, like vision, may be so tied up with the nature of the human mind and its computational properties that no design that goes against those properties will ever be more than a special purpose hack’ (McDonald, 1987). A similar methodological stance is adopted in this work, which has NLG as its domain of inquiry, with a focus on GRE. For this reason, the following chapters will contain both empirical and computational investigations into aspects of referential communicative behaviour.

1.2.1 Aspects of reference

There are various actions that could qualify as referring actions. Some of these, such as pointing, involve a non-verbal modality. Even in the verbal modality, referential expressions can take various forms. This thesis will be primarily concerned with the process that Bach (1994) has called **descriptive reference**, that is, object identification through the ascription of properties to an intended referent. One prototypical way in which such references are realised is in the form of **definite descriptions**, of which several examples were given in the previous section.

As illustrated by our hypothetical situation, the basic goal of reference is the **unambiguous identification of an intended referent**. In order for this to be possible, the target referent must be ‘available’ to the speaker in some mental representation, whether as a result of perceptual input, as in our example, or through the mediation of memory. Let us assume that there are mental ‘tokens’ representing the object(s), or **entities**, in a **domain**, and that there are finitely many **properties** of these entities, which can be used to refer to them. For instance, in our earlier example, the token in question is marked by the variable e_3 in the picture, and its properties range from the perceptual (e_3 is the leftmost but one figure), to those which are less evident from perception alone (e_3 is a princess, the other female figures are her maids, e_1 is a painter, and so on). Clearly, some further process has to be responsible for ‘translating’ these properties – which are in some sense preverbal – into words and linguistic phrases.

Pretheoretically, then, reference to an entity is a **mediated process**, since the target – an ‘object-in-the-world’ – is mentally represented by a speaker before a description is uttered. The mental representation mediates between the actual object and the linguistic expression used to identify it. Similarly, a listener uses a description to *form* a representation of the entity intended by his interlocutor, and it is on the basis of this interpretation of a speaker’s description that **resolution** – actual object identification – can take place. This model of reference as a mediated process is illustrated in Figure 1.2, in relation to the painting in Figure 1.1.

The figure displays the tasks of a speaker and a listener in a schematic form. A speaker formulates and utters a description, a listener interprets it and resolves it. Both roles, however, share two two fundamental processes. First, a representation of a referent is formed; in the case of the speaker, this is mediated by perception or recall of the relevant aspects of a domain, whereas for the listener, it is the speaker’s description that triggers the formation of this representation. The mental representation – or **conceptualisation** – in both cases consists of an entity and its properties. In both the speaker and the listener roles, **search** and **selection** are also central. For

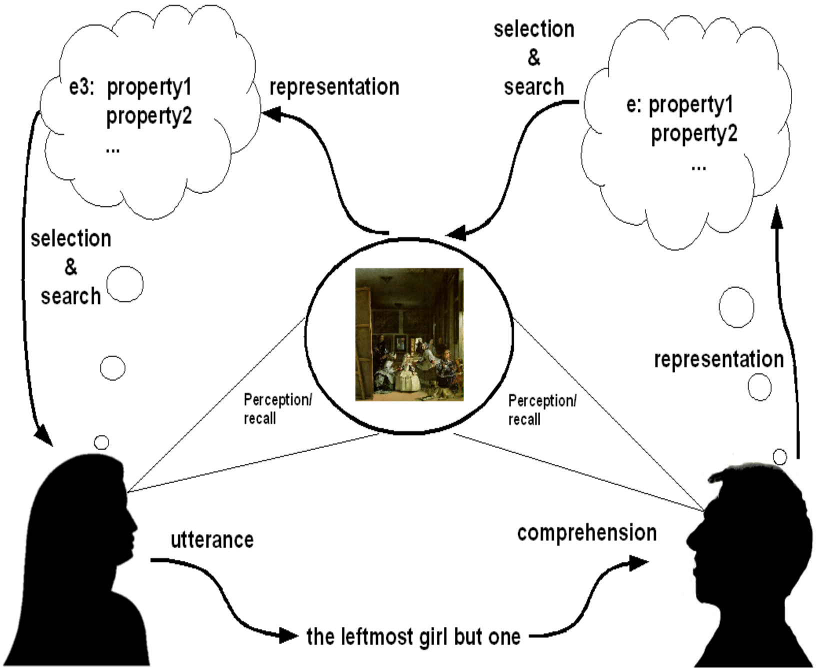


Figure 1.2: Reference as a mediated process: Tasks for speakers and listeners

the speaker, the task is to search for and select those properties of the intended referent that will help her achieve her aim (i.e. to identify the referent). The listener's search task is to find, on the basis of his representation of the speaker's message, the entity that the speaker has in mind. The different points at which these tasks are triggered results in some asymmetry in the roles of speakers and listeners. The speaker has some set of properties of an intended referent to choose from. The result of this selection, or **content determination**, is a description on the basis of which a listener resolves the reference by searching in the relevant domain.

In a sense, NLG is closest to the speaker's role (as is evident in both definitions of NLG cited above), since its focus is primarily to produce text. Indeed, GRE itself, as a subfield of NLG, has usually been defined as a content determination process involving the search for properties of an intended referent that identify it unambiguously (e.g. Dale and Reiter, 1995). Nevertheless, since a speaker's aim is to convey something to a listener, it could be argued that she should also take into account what facilitates the listener's task. As the initial example showed, there could be many such considerations, including a consideration of what knowledge is shared among the interlocutors. It seems desirable for a designer of an NLG system to take such factors into account, if the output of the system is to be 'understandable' as suggested by Reiter and Dale (1997).

My focus in this work is on the mechanisms that facilitate the conceptualisation of an intended referent. For a speaker, the conceptual representation of the referent determines what properties are available for selection and inclusion in her utterance. For a listener, the ease with which a reference is resolved will depend on how easy it is to form a 'mental picture' of the referents based on the speaker's description. If the model in the Figure is correct, this is fundamental to the human process of referring. One of the aims of this thesis is to investigate some of the principles that facilitate the representation of referents, and to use the empirical findings to inform the design of GRE algorithms. Although the investigation is not carried out within a dialogue setting, it is nevertheless important to take both the speaker's and the listener's points of view if (a) an algorithm is to serve as a viable model of a speaker in a given communicative role, and (b) its output is to be easily comprehended. Another aim of this thesis is to investigate these mechanisms in situations where the object of a referential intention is an arbitrary set, rather than a singleton or individual. To see why this is important, it's worth taking a brief tour through the main issues tackled in GRE to date.

1.2.2 A brief history of reference in NLG

If we look at work in GRE from a distance, without dealing with the specifics (this is the topic of Chapter 2), it is possible to identify two main issues whose interaction has motivated developments in the field:

1. **Adequacy**: the selection of content for a reference such that it approximates what a human speaker would do in the same situation;

2. **Efficiency**: how to achieve an adequate description without incurring the kind of computational cost that would make the entire exercise unfeasible.

As regards adequacy, several authors have motivated GRE algorithms using insights from pragmatic and semantic theory and, to a more limited extent, psycholinguistics. The theoretical backbone of most of this work in the late 1980s and early 1990s was the pragmatic theory expounded by Grice (1975). This work emphasised **brevity**. Like our hypothetical speaker in the preceding example, it required a GRE algorithm to avoid misleading the hearer by including more information than absolutely required to distinguish the referent. It turned out that under this definition, the two desiderata outlined above exhibited a tension, in that adequacy couldn't be satisfied by a polynomial-time algorithm. Things reached a climax with the publication, by Dale and Reiter (1995), of an algorithm which made two main contributions. First, it showed that content determination in GRE could be done very efficiently, if the definition of what made an adequate description was relaxed. Second, it justified the relaxation of adequacy with reference to what psycholinguists have shown that people do. Unfortunately, this work (and much later work) stopped short of actually evaluating systems empirically; moreover, the psycholinguistic motivation was, after Dale and Reiter's work, often simply held as a background assumption. As a result, once the scope of GRE algorithms was extended to deal with plurals, the problems of adequacy and efficiency again reared their heads.

Plurals are extremely understudied, both in the psycholinguistic and the computational literature. This is surprising, given that plurality is pervasive in NL discourse. For instance, a cursory survey of a sample of ca. 6000 definite descriptions in the British National Corpus revealed an approximately equal number of morphologically singular and plural descriptions. The frequency of plurals increases when we also include coordinate definite NPs such as *the princess and the maid to her right*. In principle, plurality shouldn't be a problem for GRE; it is simply a natural generalisation where an algorithm is called upon to identify an arbitrary set rather than an individual. Yet, for reasons which are somewhat reminiscent of the problems encountered by our hypothetical speaker in replying to the question *Which of the figures are the maids?*, generalising GRE in this way proved more difficult, in relation to both adequacy and efficiency. The first approaches to the problem took an exclusively logical approach (van Deemter, 2002), and actually showed that the initial assumptions of Dale and Reiter, once extended, lose the very property that had been their major selling point, namely efficiency. Moreover, the output was no longer so easily justifiable on psycholinguistic grounds as it had been when the domain of application was the relatively simple singular case. Subsequently, much of the research in the area has returned to the initial, strict Gricean definition of adequacy (often relinquishing the efficiency desideratum) (e.g. Gardent, 2002). Therefore, the recent history of GRE in a sense comes full circle, and displays an important lacuna which has hindered it from making incremental improvements.

This thesis will address this gap. In line with the methodological stance outlined in the previous section, the formal models developed in later chapters will be motivated by empirical investigations into aspects of plural reference and their relationship to singular reference.

1.3 Hypotheses

It has been known for some time that reference to singletons by human speakers (and its resolution by listeners) is influenced by a variety of processes. Of particular interest to the present work is the hypothesis, put forward by a number of psycholinguists (e.g. Pechmann, 1989), that the representation of the object of a referential intention has the structure of a **gestalt**, that is, entities are not mentally represented as bundles of separable attributes, but as conceptual wholes. Some properties are more central to this representation than others, and this affects the way speakers refer because the content of a speaker's description reflects the mental representation of the referent. Evidence for this hypothesis is discussed in the following chapter. In generalising this Gestalts Principle to sets, the main hypothesis I will investigate is that a gestalt representation also underlies reference to multiple objects, and that in this process of conceptualisation, **similarity** plays a crucial role. From a formal point of view, a description of a set is a **cover** of its elements. By hypothesis, such a cover is perceived as more unified if it conceptualises elements of the set as having something in common, whether these commonalities are perceptual attributes (such as their colour, size or shape), or are non-perceptual but inferrable from the way a set is described. If the hypothesis turns out to be correct, then GRE algorithms should aim to generate descriptions of sets whereby elements of a set are described in similar ways. In order to achieve this, the content determination process must take into account not only whether a description of a set is distinguishing, but also whether it includes properties that permit a unified conceptualisation of the set in question. This 'unified conceptualisation', or **Conceptual Coherence**, will therefore be the main theme of this work. Intuitively, we perceive a text as coherent when it 'hangs together', that is, it is well structured and its content forms a single conceptual whole. As a result of our experience of the world and of language, we often have expectations as to how things should hang together. Ultimately, the computational models proposed in the following chapters, based on empirical data, will aim to achieve conceptual coherence in reference. By hypothesis, this will result in a closer match to what speakers do, and will facilitate listeners' comprehension and resolution processes.

1.4 Outline and contributions of the thesis

The rest of this thesis is divided into two main parts. Chapters 2, 3 and 4 consist of an in-depth theoretical appraisal and empirical evaluation of existing GRE algorithms against experimentally collected data. Chapters 5 through 7 build on the empirical results of the first part to directly investigate, through corpus-based and experimental work, the predictions outlined in the previous section, using the empirical results to inform the design of GRE algorithms. Thus, while the scope of the first part is somewhat broad, it serves to raise points of empirical and theoretical interest which are taken up in the second part, where the focus is narrowed down to questions which are directly related to plurality and plural descriptions. The rest of this section outlines the structure of the thesis.

I A theoretical and empirical appraisal of the state of the art

Ch. 2 After contextualising the task of GRE within NLG as a whole (§2.2, p. 25), this chapter proceeds with a formal definition of the problem of Generating Referring Expressions (§2.4, p. 31) which is shown to cover many of the models and frameworks proposed in the literature to date and also serves as a reference point throughout the thesis. In line with much work in the area, the definition focuses on the semantic heart of the problem (content determination). The chapter then proceeds with an overview of GRE, focusing on the two main concerns outlined above, namely adequacy and efficiency. An appraisal of early models (§2.5, p. 33) segues into an exhaustive review of psycholinguistic research on reference (§2.6, p. 38), with particular reference to the Gestalts Hypothesis referred to earlier in this chapter. The focus on psycholinguistic research is in line with the empirical/cognitive stance taken throughout this work. It also serves to provide some of the background that motivated the Incremental Algorithm of Dale and Reiter (1995) and subsequent proposals that took this as a starting point (§2.7, p. 47). Among the latter, algorithms to generate plural descriptions are given particular attention (§2.7.5, p. 58). One conclusion of this chapter is that GRE as a field has suffered from a lack of empirical research, often characterised by a limited account of related work in psycholinguistics and by a lack of empirical evaluation.

Ch. 3 Chapter 3 seeks to address some of the empirical shortcomings in previous work, describing the design and annotation of the TUNA Corpus of referring expressions. Given the semantically intensive nature of the GRE task, the starting point is a recognition that such a corpus needs to be semantically transparent, in order to enable rigorous evaluation of the semantic forms that usually constitute the output of GRE. The methodology for constructing and annotating the TUNA Corpus, which meets these requirements, is laid out (§3.2, p. 67). Its emphasis on *balance* in the data implies the use of a controlled psycholinguistic experiment rather than opportunistic data collection. The annotation of the corpus (§3.5, p. 78) may also serve as an example of the kind of markup that meets the semantic transparency requirement. The rest of this chapter is dedicated to an empirical investigation of the referential descriptions produced by human authors in the corpus (§3.6, p. 83). Of particular interest are issues related to overspecification and underspecification in reference, attribute preferences as predicted by the Gestalts hypothesis (§3.8, p. 89), and the differences between singular and plural descriptions (§3.9, p. 94). The latter are a relatively understudied phenomenon in psycholinguistic research as well as in computational GRE.

Ch. 4 The corpus described in the previous chapter is now used for a speaker-oriented evaluation of some classic GRE algorithms, including the Incremental model, which compares the (semantics of) algorithm-generated and human-authored descriptions. Following an overview of previous evaluation studies in the area (§4.2, p. 103), the chapter briefly describes an implementation of the algorithms in question, compatible with the problem definition given in Chapter 2. One of the main concerns of the evaluation study itself (§4.3, p. 107) is to abstract away from differences in realisation and lexicalisation among authors (a problem with previous evaluations) and this relies heavily on the semantically transparent nature of the corpus. Another main concern, particularly in relation to the Incremental Algorithm, is to compare different incarnations of the procedure, which is shown to be highly dependent on externally set parameters (§4.5, p. 120). None of the algorithms tested has a perfect fit to the human data, though some versions of the Incremental procedure perform best. On the other hand, the algorithms are also shown to perform extremely poorly on plural data (§4.6, p. 123), when they are extended using an algorithm proposed by van Deemter (2002). This result, against the rest of the empirical background, forms the motivation for the work in Part 2.

II A psycholinguistic and computational investigation of plural reference

Ch. 5 Given the poor performance of standard GRE algorithms on plural data, this chapter begins with a more in-depth empirical investigation of the plural descriptive strategies used by authors in the TUNA Corpus, testing hypotheses that are based on Pechmann's Gestalts Principle (cf. §1.3 above). The data analysis (§5.2, p. 130) finds three important properties of plural descriptions: (a) authors tend to partition sets according to the basic-level TYPE of their elements; (b) perceptual properties which are highly central to the mental representation of a referent are often included in such *partitioned descriptions*; (c) elements of a partition evince *semantic parallelism*, that is, they are often described using the same attributes, even when these are redundant, though this depends crucially on the centrality of such properties to the Gestalt representation of an object. The findings inform the design of a new algorithm for referring to sets of arbitrary size (§5.3, p. 140). Unlike previous algorithms, this one separates content determination proper from the construction of a linguistically transparent logical form. The latter is based on an on-the-fly partitioning strategy. Content determination incorporates a corpus-derived statistical model to determine whether a property should be used to describe each element of a partition, even when it is not required for identification. An evaluation (§5.4, p. 150) shows that the new algorithm performs significantly better on the plural data in the TUNA Corpus than a previous model. Chapter 5 takes these results further by considering strategies for aggregation in plural descriptions, partly motivated by the possibility that partitioning may be less in evidence in case elements of a referent have the same basic-level TYPE. A new corpus study is described, which focuses on semantic constraints and complexity limitations on aggregation in plurals (§5.5, p. 154). The results inform the design of an aggregation algorithm, whose integration with the partitioning-based content determination procedure is also discussed (§5.5.4, p. 161).

Ch. 6 While the previous chapter showed that perceptual similarity and semantic parallelism exerts an influence on people's plural descriptive strategies, the Conceptual Coherence Hypothesis is extended and generalised here to descriptions of pluralities in discourse. The main focus is on cases where several possible ways of describing (and hence conceptualising) a set of referents exist (for instance *the professor and the Italian* versus *the professor and the lecturer*). The hypothesis tested in this chapter, which has some precedent in previous psycholinguistic and formal semantic work (§6.2, p. 167), is that descriptions are better if the choice of properties for different elements of a set is *semantically similar*. Various formal definitions of semantic similarity (§6.4, p. 176) are tested in an initial series of three experiments, the most adequate of which is found to be a corpus-based, distributional definition based on word occurrence in the same grammatical contexts (§6.5, p. 178). These initial experiments use *magnitude estimation*, a method for eliciting ratings of stimuli, and provide validation data that shows that this method is feasible in a study of this kind. They are followed by two experiments on similarity constraints on the way people produce plural references. The first of these shows that the hypothesis correctly predicts that people are more likely to generate a plural description when the elements of a plurality can be conceptualised in similar ways (§6.6, p. 192); the second investigates content determination in plural reference, showing that given a choice of more than one way of describing a set, similarity exerts a strong influence on people's choices. The hypothesis that these results support can be taken to characterise a family of GRE algorithms, those which seek the most conceptually coherent description (where this notion is interpreted as 'the description under which a set is covered using the most similar properties available'). These aims are different from those proposed in previous GRE work on plurals. A sixth experiment therefore compares the predictions of one class of such models, those emphasising brevity and conciseness, to the Conceptual Coherence model, finding no evidence in favour of the former, but strong evidence for the latter (§6.8, p. 200).

Ch. 7 This chapter ports the results of the psycholinguistic experiments of Chapter 6 to the GRE domain. Because the definition of similarity that was found to be most adequate was distributional, and holds between *words* rather than *properties*, the first step is to define the notion of a *lexical item* as a pairing between a word-form and a semantic representation. This leads to a graph-based definition of a lexicon in which lexical items are connected by edges whose weights reflect the semantic similarity between them (§7.2, p. 209). The lexicon constitutes the basic input to a content determination procedure, making lexicalisation part and parcel of content determination. Two such procedures are described, both of which are presented as possible instantiations of the Conceptual Coherence model. One of them (§7.6, p. 226) precedes content determination by a procedure which clusters together related words into 'conceptual perspectives', broadly conceived as sets of words which are used in similar contexts and conceptualise things in related ways. The second model (§7.7, p. 230) is based on a view of the lexicon as an active repository of information, whereby the use of a lexical item (its selection by an algorithm) results in spreading activation to nearby items, making them easier to retrieve. These two models differ in that, while the first attempts to precompile the available 'conceptual covers' for a set in clusters of related words, the second is explicitly priming-based, and therefore models the phenomena found in Chapter 6 using a more basic mechanism. These and other theoretical differences between the two models are discussed in some detail (§7.8, p. 233). It should be noted, however, that the two algorithms are intended as possible instantiations of the family of algorithms under the Conceptual Coherence model, rather than as an exhaustive coverage of the space of possible instantiations.

The journey outlined above, from the current state of the art in GRE to an investigation of the role of perceptual similarity and conceptual coherence in reference, is synthesised in the final chapter. As the preceding outline might suggest, the present work represents an attempt to tread the line between cognitive (psycholinguistic) theories and methodologies, and algorithmic application, based on the view that computational solutions to cognitive problems are best approached from both angles.

Chapter 2

Generation of Referring Expressions

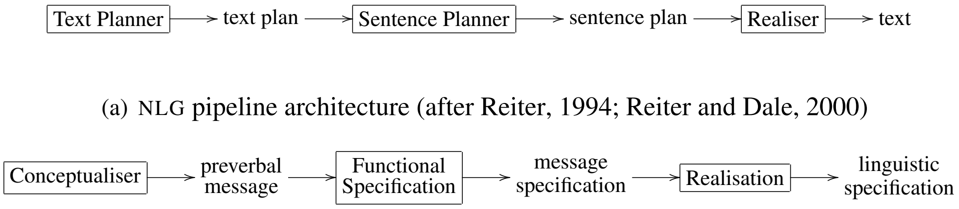
2.1 Introduction

This chapter provides a detailed review of the state of the art in the Generation of Referring Expressions. After contextualising the task of GRE within Natural Language Generation (NLG) Systems (§2.2), it introduces some of the fundamental assumptions made in the GRE literature (§2.3, p. 28). This gives rise to a problem definition (§2.4, p. 31), which formally defines the basic components of a GRE *problem instance* in a declarative fashion. This definition, which serves as a reference point throughout the remainder of the thesis, is shown in this chapter to apply to many of the best-known algorithms proposed in the field.

Focusing particularly on a family of algorithms proposed by Dale and Reiter, the discussion then proceeds as follows. I begin with some classic models whose definition of referential adequacy was inspired by the Gricean maxims of communication (§2.5, p. 33). Later work raised concerns both about this definition of adequacy and about the computational tractability of these procedures. In a somewhat parallel fashion, work in psycholinguistics on the production and comprehension of referring expressions was also questioning the predictive validity of models based on a strict interpretation of the Gricean maxims. Therefore, a substantial section of this chapter (§2.6, p. 38) is dedicated to this body of work. This serves as some of the backdrop against which the Incremental Algorithm, the gold standard in the field, was proposed (Dale and Reiter, 1995) (§2.7, p. 47), though it should be emphasised that the interface between psycholinguistic and computational research in this area has been tenuous and opportunistic. As the discussion of later models which build directly on the Incremental Algorithm shows, one characteristic of the field of GRE has been an insufficient acknowledgement of the relevant psycholinguistic literature, and a tendency to stop short of evaluating algorithms empirically. I raise these issues at various points, including the discussion of context-sensitive GRE (§2.7.1, p. 50), relations (§2.7.3, p. 54) and gradable properties (§2.7.4, p. 55). However, this point can be made even more forcefully in relation to algorithms for the generation of plural references which, as hinted at in the previous chapter, have occasionally signalled a return to the 'strict Gricean' interpretation of adequacy that was questioned in psycholinguistics over several decades of research.

2.2 The place of Referring Expressions Generation in NLG systems

The task of Natural Language Generation (NLG) systems is to produce natural language (NL) text from an underlying input representation, based on a Knowledge Base. The nature and representation of the inputs depends on the application domain, while the textual output is often the result of



(a) NLG pipeline architecture (after Reiter, 1994; Reiter and Dale, 2000)

(b) Psycholinguistic production architecture (after Levelt, 1989)

Figure 2.1: Pipeline architectures in psycholinguistics and NLG

a number of processing stages. Surveys of NLG in Reiter (1994) and Reiter and Dale (2000) have suggested that many NLG systems conform to a basic tripartite architecture, whose components are organised in a pipeline which, as shown in Figure 2.1(a), consists of the following three stages:

1. **Text planning**: The formulation of a message at an abstract level, specifying the communicative intention(s) to be achieved and the pragmatic (sub-) goals of the component communicative acts, and mapping these goals to utterance-level segments, or messages.
2. **Microplanning (sentence planning)**: The fleshing out of the content of these messages in greater detail, through:
   (a) *Generation of referring expressions* (GRE): The selection of content for what will eventually become the noun phrases denoting domain entities;
   (b) *Lexicalisation*: Selection of appropriate lexemes to express the predicates in the message;
   (c) *Aggregation*: Aggregation of multiple message fragments into more cohesive units
3. **Realisation**: The mapping of the semantic representations generated in previous stages onto their natural language representation, by applying language-specific syntactic and morphological rules.

The architecture highlights widely-held distinctions between different sub-problems of NLG. Indeed, further detailed surveys of existing NLG systems have suggested that while researchers may disagree about the precise location of some task in the overall architecture of a system, a number of such tasks are almost universally held to be necessary, and their scope is also well agreed-upon (Paiva, 1998; Cahill et al., 1999; Mellish et al., 2006). The general picture presented in the Figure is that of a process which, starting from a communicative intention, goes through successive stages, each of which is **computationally autonomous** and **informationally encapsulated**, to result in a realisation of that intention via a NL utterance. While the output of one stage constitutes the input to the next, the internal workings of any particular component or module is largely independent of any of the others'. This echoes the Fodorian Modularity of Mind thesis (Fodor, 1983), which rests on a binary distinction between general purpose higher-order cognitive processes such as reasoning, and lower-level cognitive functions such as perception and natural language production and understanding. The latter functions are assumed to be modular and independent. The binary distinction is reflected to some extent in the NLG literature. Until the late 1980s, generation was divided into a **strategic** and a **tactical** component (e.g. Appelt, 1987a), a distinction that is still evident in the division of labour of the more recent tripartite architecture. While tactical generation is responsible for the 'strictly linguistic' task of realising messages using language-specific morphosyntactic rules, the term 'strategic generation' subsumes all the pre-linguistic reasoning processes that are involved in the construction of an utterance. The problem space of strategic generation is populated by semantic objects, specified in a formal language; its output is a semantic representation that is mapped to a NL utterance by the tactical component.

The finer-grained distinctions introduced by the tripartite architecture sometimes blur the boundaries between purely 'semantic' and strictly 'linguistic' tasks. This is especially true of microplanning. For instance, much of the literature on Generation of Referring Expressions (GRE) focuses on the semantic task of Content Determination, whose search space is populated by properties – i.e. *semantic* objects. However, microplanning also consists of sub-tasks – namely lexicalisation and aggregation – that bring the semantic specification of a message closer to its eventual linguistic realisation. Reiter and Dale (2000) have suggested (albeit with reservations) that these three sub-components of microplanning should themselves be organised in a pipeline. As we shall see, however, recent work in NLG has questioned the strong modularity implied by this model, arguing for a more interleaved architecture that couples semantico-pragmatic and morphosyntactic processes more closely. As an example, the SPUD microplanning system plans utterances by incrementally constructing representations which combine lexico-semantic, pragmatic and syntactic elements, keeping track of the contribution of new linguistic material added to an utterance to the realisation of a communicative intention (Stone et al., 2003). As we shall see (§2.7.7, p. 63), the interleaving of syntax, pragmatics and semantics has also been a feature of recent work in GRE. This thesis will also propose some revision to the separation of microplanning tasks, especially with regard to dependencies between content determination and lexicalisation.

Reiter (1994) hypothesised that the tripartite division of NLG tasks could serve as a model of the psycholinguistic processes underlying language production. As a comparison of Figures 2.1(a) and 2.1(b) reveals, there is in fact a parallel between the NLG architecture and the dominant model of the human production system proposed in the psycholinguistic literature (Levelt, 1989; Bock and Levelt, 1994). In a manner that recalls the strategic/tactical distinction, the latter also distinguishes between **message-level** or **conceptualisation** processes that deal with the planning of utterances, and linguistic processes of **grammatical encoding**. However, there are important differences between the two. In particular, content determination for referring expressions, part of conceptualisation (macroplanning), is separate from lexicalisation, which, like aggregation (Kempen and Hoenkamp, 1987), is part of grammatical encoding. Nevertheless, there is a striking amount of agreement between the two models, in terms of the specific sub-tasks involved in the production of a linguistic message.

One of the reasons why a pipeline architecture is attractive from a psycholinguistic point of view is the **incrementality** of human language production. This processing characteristic, which Levelt has termed **Wundt's Principle**, implies that *each processing component will be triggered into activity by a minimal amount of its characteristic input* (Levelt, 1989, p.26). In other words,

	TYPE	ROLE	CLOTHING	POSITION	POSTURE
e_1	man	painter	wears_black	left-of(e_2)	standing
e_2	woman	maid	wears_grey	left-of(e_3)	kneeling
e_3	woman	princess	wears_white	left-of(e_4)	standing
e_4	woman	maid	wears_brown	left-of(e_5)	standing
e_5	woman	maid	wears_black	left-of(e_6)	standing
e_6	girl	–	wears_red	right-of(e_5)	standing
e_7	dog	–	–	front	sitting

Table 2.1: Simplified representation of Figure 1.1

the production of a linguistic message at any stage of the process need not await the completion of the previous process. Rather, the message begins to be formulated as soon as sufficient conceptual material has been assembled and the articulation of a message (the final stage of Levelt's model, not shown in the figure) can proceed as the message is constructed. This characteristic of human language production can be captured with minimal alterations to the computational generation architecture. For instance, Guhe et al.'s (2004) Incremental Conceptualiser generates conceptual structures that feed into microplanning and realisation modules at the earliest possible point. The adherence to Wundt's principle allows the system to describe dynamic scenes as they unfold in time.

2.3 GRE: Models and Frameworks

Since the foundational work of Appelt (1985a) and Dale (1989), Generation of Referring Expressions (GRE) has been the focus of extensive research. This has resulted in a significant consensus emerging over the basic problem definition, its inputs, and its output. This section focuses on these aspects, deferring detailed discussion of different approaches to GRE to later sections.

The 'consensus' architecture of Reiter and Dale places GRE in the microplanning stage of NLG systems, at which point the abstract structure of a message has been formulated. Such message structures are assumed to contain non-linguistic identifiers for domain entities, and the task of GRE is to select the content for noun phrases (NPs) that identify these entities for a listener or reader. A related task is to take into account discourse context to determine the form of the NP produced, whether it is to be realised as a full noun phrase or as a pronominal anaphor (Kibble, 1999; Reiter and Dale, 2000; Krahmer and Theune, 2002). Content determination is achieved by searching through a Knowledge Base (KB) containing a finite set of entities, each with finitely many properties, often specified as attribute-value pairs (Dale and Reiter, 1995). For example, Table 2.1 represents a fragment of the 'domain' in the painting that constituted the motivating example in the previous chapter. The top row of the table specifies a list of attributes; each row corresponds to a domain entity, with cells in the table being the attribute-values of that entity.

In line with the example from the previous chapter, let us assume that a message has been formulated in answer to the question *Which of the figures in the painting is the Infanta?*. A satisfactory answer to this question requires the unambiguous identification of a domain entity, in this case e_3. Correspondingly, the dominant assumption of most work in GRE following Dale (1989) has been that the task of GRE algorithms is to produce **distinguishing descriptions**, typically realised as definite NPs, such as (2.1–2.4) below. The communicative situation described in §1.1 has a further characteristic that has achieved currency in the literature, namely that the domain of discourse is **mutually known** to the interlocutors (the painting is visible to both speaker and hearer).

(2.1) the princess

(2.2) the third figure from the left

(2.3) the person to the right of the kneeling woman

(2.4) the girl in the white dress

The notion of a distinguishing description has its roots in a tradition in philosophy and formal semantics which starts from the work of Frege (1952) and Russell (1905). Russell observed that while an indefinite NP presupposes the existence of the entity denoted, a definite description requires not only that the referent exist, but that the properties predicated of the referent hold of that referent alone. Since Strawson (1950), the existence and uniqueness properties of referring NPs have often been viewed as pragmatic presuppositions, rather than semantic rules of interpretation. Strawson argued that it is not the description itself which is referential, but it is the speaker that refers *via* a description. Hence, reference is an intentional act on the part of a rational agent, and a description is a realisation of the speaker's referential intention.

One source of evidence in support of the pragmatic view is the observation that sometimes references do not successfully distinguish a referent on the basis of their semantics, yet interlocutors still manage to successfully identify the referent, suggesting that they are aware of the speaker's referential intentions (cf. Donnellan, 1966). In line with this view, Searle (1969) distinguished between the speaker's *intention* to identify, and her *ability* to do so successfully. His account of reference, subsumed under a general theory of Speech Acts, is based on a set of pragmatic rules of communicative behaviour, in which the speaker's intentions occupy centre-stage. These intentions are twofold: (a) to identify a referent in a domain for a hearer, and (b) to get the hearer to recognise the speaker's intention to identify the referent, given the hearer's knowledge of the rules governing referential acts. Searle also emphasised the role of context in a referential act, since it is only with respect to a mutually known (or mutually accessible) domain of discourse that a hearer can successfully establish which referent is intended.

These two aspects of the referential communicative situation have been emphasised to different extents in the GRE literature. One line of research, following Appelt and Kronfeld (Appelt, 1985a,b, 1987b; Appelt and Kronfeld, 1987) has developed computational models of reference strongly influenced by Speech Act Theory. In this framework, the generator is viewed as a model of a rational agent, whose referring actions are the outcome of a planning process aimed at aligning the putative hearer's beliefs about an intended referent with the agent's and, once this is satisfied, to enable the hearer to identify the referent. For instance, Appelt (1985a,b) proposes an intensional logic of beliefs and actions, within which the preconditions of a successful referring action can be defined. The generator plans its actions accordingly. A crucial component of this process is the extent to which knowledge about the domain and the intended referent is shared between the generator and the hearer.

Mutual knowledge implies that a speaker must reason not only about what a hearer knows, but also about the fact that the speaker knows that the hearer knows this, and that the reverse also holds. As Clark and Marshall (1981) pointed out, this can lead to an infinitely recursive reasoning process of the form *S(peaker) knows that H(earer) knows that S knows ....* The reason why this doesn't happen, they proposed, was that heuristics are routinely used by interlocutors, who rely on different sources of shared knowledge and are cognisant of the preconditions for successful reference. Thus, if a property p is believed to be true of a referent r, and the speaker believes the hearer knows this, then the mutual knowledge condition is assumed to hold. In Appelt's (1985a) KAMP system, some reasoning along the same lines is performed in the process of planning an NP.

These models subsume reference under a more general theory of communicative action in a joint setting, a position reminiscent of the *language-as-action* view of Clark (1996), which makes the role of mutual knowledge central to the referential process. Clark distinguished this from a *language-as-product* view (primarily in the psycholinguistic literature). *Language as action* underlines the collaborative process of communication; hence, the primary arena for the investigation of such processes is dialogue (Clark, 1997b), and the primary focus is on the rational processing underlying communication. The *language as product* tradition emphasises the mechanistic processes involved in the production or comprehension of utterances, and has often studied them in non-dialogic contexts (but cf. Pickering and Garrod, 2004).

Models of reference in this paradigm have proven particularly influential in the GRE literature on dialogue (e.g. Edmonds, 1994; Heeman and Hirst, 1995). Dialogue-based models go beyond Appelt's framework of reasoning about a hearer's beliefs, viewing the alignment of speaker and hearer models as a process of negotiation between interlocutors, along the lines proposed by Clark and his colleagues (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996).

A rather different line of research has emerged following the work of Dale and Reiter (Dale, 1989; Reiter, 1990; Dale and Reiter, 1992, 1995). Whereas Appelt's model was intended as a broad framework to encompass different uses of noun phrases in natural language, with reference as a special case, the focus in the Dale/Reiter line of research has been on Content Determination proper. In terms of the model in Figure 1.2 (p. 18), this is the part of the referring process that deals with a speaker's search and selection, thereby fleshing out the content of a referential NP. Rather than developing explicit models of collaborative communication and the establishment of common ground, work in this paradigm has tended to hold mutual knowledge of the domain as a background assumption, and has been primarily motivated by the need to develop an explicit and computationally tractable definition of what it means for a description to be distinguishing, and what properties of a description make it **adequate**, given that its function is primarily to identify. Gricean principles of cooperative communication have played the central role here (Grice, 1975): since the principal communicative aim is identification, these models have often defined an adequate description as one which does not communicate information that is not necessary to achieve this goal, as this gives rise to false implicatures in the hearer.

2.4 Problem definition

Having reviewed some of the basic underlying assumptions in GRE, we are now in a better position to give an explicit problem definition, focusing on GRE as Content Determination. I will start by defining the properties of a GRE *problem instance*, and the success criterion of a GRE algorithm. These definitions are formalisations of the notions found in Dale and Reiter (1995). The discussion then turns to some of the consequences of these definitions. Since a Knowledge Base (KB) is the primary ingredient of an GRE problem instance, I begin with a simple definition of a KB.

Definition 1. Knowledge Base
A KB is a double $\langle U, \mathbb{P} \rangle$ where:

- $U = \{e_1, \ldots, e_n\}$ is a denumerable set of domain entities (the 'universe of discourse');
- $\mathbb{P} = \{\langle \text{A} : v \rangle | \llbracket \langle \text{A} : v \rangle \rrbracket \subseteq U\}$ is a set of properties (finitely many), represented as attribute-value pairs.

Throughout this thesis, I will maintain this notation, using uppercase letters to denote attributes, italicising names of values of those attributes. Moreover, I will switch between two possible 'views' of a property – as a literal p, and as an attribute-value pair $\langle \text{A} : v \rangle$, depending on the requirements of the discussion. Thus, $\langle \text{TYPE} : \textit{man} \rangle$ is a property, and $\llbracket \langle \text{TYPE} : \textit{man} \rangle \rrbracket = \{e_1\}$ in Table 2.1.

A core assumption of virtually all the algorithms discussed below is that the KB contains all and only the relevant knowledge for the task. This Closed World Assumption has some computationally attractive consequences, in that it simplifies the task of performing inferences about entities on the basis of the available knowledge, and also ensures that $\mathbb{P}$ is finite.

Definition 2. GRE Problem Instance
A GRE problem instance is a 4-tuple $\langle K, r, P_r, D \rangle$ where:

- $K = \langle U, \mathbb{P} \rangle$ is a KB;
- $r \in U$ is an intended referent;
- $P_r \subseteq \mathbb{P} = \{p \mid r \in \llbracket p \rrbracket\}$ (The set of relevant properties, i.e. those true of r);
- $D \subseteq P_r$ is a description of r

Definition 3. GRE Success Criterion
A GRE algorithm is successful iff:
$\llbracket D \rrbracket = \bigcap_{p \in D} \llbracket p \rrbracket = \{r\}$

According to the above definitions, the primary ingredients of a GRE problem instance are the domain entities and the properties that are known to hold of them. Success is a matter of finding a set of properties which uniquely distinguish a referent.

Three aspects of the problem are worth emphasising, in view of the arguments presented in the following chapters. First, the definition assumes (essentially following Dale and Reiter) that there is one intended referent, that is, all descriptions generated are singular. Generalising the problem definition to plurals would require some revision, since reference to sets will not be guaranteed if success is defined exclusively with respect to set intersection (van Deemter, 2002). Second, GRE is defined as a **semantic problem**: it is KB properties that populate the search space for a GRE algorithm, and the description generated is simply defined as a subset of the properties true of the referent. Third, the success criterion is *extensional*: if r is the only referent of D, then the algorithm can be said to have been successful. Generalisations of GRE to plurals are discussed in the following section. Here, I turn to the consequences of the 'semantic bias' of Definition 2 and the extensional character of Definition 3.

2.4.1 Extensional equivalence and adequacy

How is a GRE algorithm to decide between coextensive alternatives, that is, different sets of properties all of which satisfy the criterion in Definition 3? Answers to this question depend on a theory of **descriptive adequacy**, whereby the 'best description of r' can be defined. This, in turn, determines how an algorithm searches for properties in P_r to find that description. Two kinds of orderings are therefore defined for a GRE algorithm which instantiates Definition 2. The first is an ordering between descriptions. Let D_r be the set of distinguishing descriptions for r. Further, let the notion of descriptive adequacy be abbreviated by $>>_{D_x}$, so that $D >>_{D_x} D'$ is shorthand for 'D is more adequate than D' by criterion x' (cf. Reiter, 1990). This yields a partial order $\langle D_r, >>_{D_x} \rangle$, and the role of a GRE algorithm is to find the best alternative which satisfies the success criterion. This, however, says nothing about the *procedure* for achieving this. It is not descriptions, but properties (or combinations thereof) that populate the search space of a GRE algorithm[1], so that search needs to be guided by some heuristic, based on $>>_x$, that orders properties with respect to each other. Let $p >>_{p_x} q$ abbreviate 'property p is a better candidate than property q for inclusion in a description, given the current state of an algorithm'. Such a heuristic can be viewed as the inverse of a **cost function**, and represents the way an algorithm might go about approximating or achieving $>>_{D_x}$.

Cost functions are the fundamental ingredient of heuristic search problems in AI (Russell and Norvig, 2003). Indeed, it has been argued that this framework offers a unified way of conceptualising content determination algorithms for GRE Bohnet and Dale (2005). The first explicit attempt to formalise several GRE algorithms within a unified search framework was that of Krahmer et al. (2003), who adopted a graph-theoretic perspective, which formalises the KB as a *scene graph*, whose vertices are the domain entities, and whose edges correspond to properties of those entities. Properties which are semantically one-place predicates are realised as loops, while 2-place relations connect the vertices representing their arguments.[2] Within this framework, GRE becomes a subgraph isomorphism problem: an intended referent is a node, and a description is a subgraph of the scene graph which contains exactly that node. This framework is attractive because it does not make a separation between the knowledge representation component and the descriptive component of the GRE problem. It also makes explicit the notions of cost and adequacy: by representing the KB as a weighted graph, the 'least-cost' alternative is explicitly available in the representation, as the distinguishing sub-graph whose total weight is minimal.

[1] Another way of putting this is that the search space of these algorithms is populated by *partial descriptions*, since any property or property combination that is true of the intended referent can be viewed as part of a potential description of that referent.

[2] Thus, a scene graph is a pseudo-graph, since more than one edge between two vertices is allowable, as are loops.

The approach I adopt here to describe four highly influential content determination algorithms proposed by Dale and Reiter will roughly follow that of Bohnet and Dale. Under this view (essentially that of Russell and Norvig, 2003), content determination algorithms undergo a succession of states, having at their disposal a dynamic **queue**, which imposes an ordering (corresponding to $>>_{p_x}$) among (combinations of) properties in the search space, and determines, at each successive state, which property is to be considered next. The simplest way to conceive of the dynamic queue is as a priority queue which holds properties or combinations thereof and maintains, at a given state of the algorithm, an ordering among combinations of properties corresponding to the transitive closure of $>>_{p_x}$. The queue has two associated functions:

- $dequeue(Q)$ returns the element with the highest priority (lowest cost) in Q;
- $enqueue(Q, p)$ places an element p in the queue, whose position relative to existing elements is determined by $>>_{p_x}$.

Under this framework, the basic structure of a content-determination procedure which satisfies Definitions 2 and 3 is as follows:

```
Require: $r$, $P_r$, $U$
initialise $C$ with the set of distractors $U - \{r\}$
enqueue all properties in $P_r$
while $Q$ is non-empty do
    dequeue the next property (the one with lowest cost)
    if the current property removes distractors then
        update the description or return the current property
    else update the queue with combinations involving the current property
    end if
end while
```

The four algorithms discussed below are primarily distinguished by how they enqueue properties, and by whether they need to search through *combinations* of properties, or only through the set of literals P_r.

2.5 GRE algorithms: The role of Gricean Brevity

In the search algorithms proposed by Dale and Reiter, the Gricean maxims of conversation are central, especially the Maxim of Quantity, which has the following two components:

1. Make your contribution as informative as required.
2. Make your contribution no more informative than required.

This maxim was interpreted by Dale (1989) as a constraint on avoidance of **overspecification**, an interpretation that was already prefigured in Appelt (1985a), and in some experimental psycholinguistic work (e.g. Olson, 1970; Ford and Olson, 1975; Whitehurst and Sonnenschein, 1978; Sonnenschein, 1982). Olson (1970) observed that reference has a primarily contrastive function, and the content of a description is determined by the **contrast set** or **distractors** from which the referent is distinguished. Kronfeld (1989) referred to this as the **functional relevance** of a description. It follows – in a Gricean vein – that any semantic content which does not contribute to a description's functional relevance is redundant and likely to give rise to a false implicature, since the reader expects the generator to be cooperative, and include only (functionally) relevant information. Along similar lines, Appelt (1985b) suggested that an agent's plan to refer should be subsumed under a more general plan to *inform*. In this view, if the sole aim is to get the hearer to identify a referent, then the generator should include no more information than necessary for this purpose; however, extra properties are justified when there is an additional intention to inform the hearer that they hold true of the referent (cf. O'Donnell et al., 1998, for a related view).

The Gricean account is based on a symmetric model of communication, in which speakers, being themselves listeners, use the same heuristics as their interlocutors, and therefore produce utterances in line with hearer expectations (Oberlander, 1998). The same holds of listeners: If false implicatures arise from the use of information over and above what is strictly necessary to identify, this is because a hearer, having understood a speaker's intention, should expect the right amount of information, and no more than that.

Psycholinguists, following Olson, viewed the process of referent identification and reference resolution as requiring a comparison between the intended referent and the distractors, in order to find distinguishing properties (cf. Figure 1.2, p. 18). This is often framed as a decision problem, whereby the speaker (and, conversely, the listener) has to decide, for a given property, whether it is functionally relevant (e.g. Sonnenschein, 1982; Deutsch and Pechmann, 1982; Belke and Meyer, 2002). This view also dominated the work of Dale (1989), who interpreted the Maxim of Quantity as a directive, proposing an algorithm that tests combinations of properties true of the intended referent in order of their length, and terminating when a distinguishing combination is found or the search space is exhausted. Thus, his **Full Brevity** (FB) algorithm defines descriptive adequacy in terms of brevity:

$$D >>_{D_{\text{FB}}} D' \leftrightarrow [\![D]\!] = [\![D']\!] \wedge |D| < |D'| \tag{2.5}$$

.

Correspondingly, the ordering relation among properties in the queue, $>>_{p_{\text{FB}}}$, prioritises shorter combinations before longer ones. Effectively, this makes $>>_{p_{\text{FB}}}$ coincide with $>>_{D_{FB}}$. To achieve this, the algorithm must maintain in the queue not only the literals in P_r, but also combinations of those literals, corresponding to logical conjunctions. At any stage in the algorithm, the function $dequeue(Q)$ is defined as follows:

$$dequeue(Q) =_{def} \arg\min_{P \in Q} |P| \tag{2.6}$$

Pseudocode for this algorithm is shown in Algorithm 1. It proceeds largely as described in the previous section, first initialising the set of distractors C [1.1] and enqueing all literals in P_r [1.3]. When a property or combination is dequeued [1.6], the test at [1.7] is for whether it uniquely distinguishes r, that is, removes *all* the distractors from C, in which case it is returned. If not, the current property (or combination) is conjoined to all other properties in P_r, and the resulting combinations enqueued [1.10].

Algorithm 1 Full Brevity Algorithm

```
Require: $r, P_r, U$
$C \leftarrow U - \{r\}$
for $p \in P_r$ do
    $enqueue(p, Q)$
end for
while $Q \neq \emptyset$ do
    $p \leftarrow dequeue(Q)$
    if $[\![ \ p \ ]\!] - C = \emptyset$ then return $p$
    else
        for $q \in P_r - \{p\}$ do
            $enqueue(p \wedge q, Q)$
        end for
    end if
end while
return $\emptyset$
```

The hallmark of this algorithm is that it does not *construct* a description by adding properties to it; rather it searches exhaustively through combinations of increasing length until a distinguishing description is found. This gives it exponential worst-case complexity. Reiter (1990) showed that there exists a polynomial-time transformation of FB to a Minimal Set Cover Problem.[3] Let $C = U - \{r\}$ be the distractor or contrast set, $p \in P_r$, and *rulesOut*$(p) = C - [\![\ p \ ]\!]$. Recall that D_r was defined as the set of distinguishing descriptions of r. Thus, it can be redefined as follows:

$$D_r = \left\{ D | \bigcup_{p \in D} rulesOut(p) = C \right\} \tag{2.7}$$

Now, FB seeks the description which satisfies the following:

$$D_{\text{FB}} = \underset{D \in D_r}{\arg\min} \, |D| \tag{2.8}$$

which is equivalent to the minimal set of properties which by (2.7) covers C. Reiter (1990) proposed an alternative approach to achieving brevity, termed the **Local Brevity** (LB) heuristic. This is defined as the transitive closure of the ordering relation $>>_{D_{LB}}$, shown in (2.9).

$$D >>_{D_{LB}} D' \leftrightarrow [\![\ D \ ]\!] = [\![\ D' \ ]\!] \wedge |D'| - |D| = 1 \tag{2.9}$$

that is, a description D is more adequate than D' if D has at least one 'component'[4] less than D', and both are coextensive. LB was conceived as a post-edit strategy to replace unnecessary components in a generated description. For instance, one could imagine a generator randomly adding properties to a description until r is distinguished, and then testing, for each property $p \in P_r - D$, whether there is a combination of properties in D that p can replace, while still

[3] Thus, no polynomial time algorithmic solution exists for FB, unless P=NP.

[4] Reiter (1990) proposes various interpretations of the term 'component', among them that it be equated with 'property' and/or 'lexical item'.

Algorithm 2 Greedy Algorithm

```
Require: $r, P_r, U$
$D \leftarrow \emptyset$
$C \leftarrow U - \{r\}$
for $p \in P_r$ do
    $enqueue(p, Q)$
end for
while $Q \neq \emptyset$ do
    if $C = \emptyset$ then
        return D
    end if
    $p \leftarrow dequeue(Q)$
    if $[\![ p ]\!] - C \neq \emptyset$ then
        $D \leftarrow D \cup \{p\}$
        $C \leftarrow C \cap [\![ p ]\!]$
    end if
end while
return $D$
```

retaining the distinguishing character of D.

Yet another tractable alternative to FB is the **Greedy Algorithm** (GR; Dale, 1989). Unlike FB, GR does not search exhaustively through all possible combinations of properties until r is distinguished. Instead, the algorithm loops through P_r, adding properties which remove some distractors to a set D (the description), and updating the set of distractors accordingly. The property selected at any point is the one with the greatest **discriminatory power**, defined as follows:

$$disc(p) = |C - [\![p]\!]| \tag{2.10}$$

The adequacy relation among alternative descriptions in GR is the same as in FB, since the aim is to produce brief descriptions. However, unlike FB, for which $>>_{D_x}$ and $>>_{p_x}$ coincide (because the algorithm performs exhaustive search), GR adds properties to a description incrementally, obviating the need to enqueue combinations of properties. Rather, the queue only maintains literals in the order defined below:

$$p >>_{p_{\text{GR}}} p' \leftrightarrow disc(p) > disc(p') \tag{2.11}$$

As a result, the function $dequeue(Q)$ for GR is defined, at any state reached by the algorithm, as:

$$dequeue(Q) =_{def} \arg\max_{p \in Q} disc(p) \tag{2.12}$$

Pseudocode for GR is given in Algorithm 2. At any stage, the property considered by GR is the one with the greatest discriminatory power, which depends on the elements of the context set which have been excluded so far. This means that every time a property is found to exclude some distractors, the algorithm needs to update C, the set of distractors [2.13], as well as the description [2.12]. Moreover, the ordering among properties in the queue changes because their discriminatory power depends on C (hence, the priority queue must be dynamic). The procedure terminates as soon as the distractor set is found to be empty [2.7]. Because it only searches through literals (unlike FB, it never enqueues conjunctions), GR is polynomial in the number of properties in the KB and the number of properties in the resulting description. Let $n_d = |D|$ and $n_p = |P_r|$. Then the algorithm tests at most n_p properties, comparing them for their discriminatory power, at most n_d times, giving it a complexity $O(n_d n_p)$.

From a historical perspective, GR is important because some of its core properties would later be incorporated into the 'gold standard' content determination procedure, the Incremental Algorithm (IA) of Dale and Reiter (1995). First, GR was motivated by a tension between a definition of descriptive adequacy and the desirability of a tractable solution to the GRE problem. As we shall see, the same argument was used to motivate the IA. Secondly, GR already incorporates a notion of **incrementality**, because it constructs a description rather than performing exhaustive search.

Although GR incorporates a definition of descriptive adequacy identical to that of FB (see 2.5), its greedy search heuristic does not guarantee that the output will in fact satisfy (2.5). The extent to which GR approximates FB depends on the size of the minimal description available for r in the KB. In case the description returned by FB is of length 1 (i.e. contains a single literal), GR and FB coincide, or at least output descriptions of identical length, since the property in the description returned by FB (or one coextensive with it) is by definition the one with the highest discriminatory power at the beginning of the loop in Algorithm 2. For example, both FB and GR might describe e_3 as *the princess*, given the domain in Table 2.1. However, there is no guarantee that the output of GR will be minimal in case the minimal description is of length two or more.[5] Consider the case where the minimal description is of length 2 (a conjunction of 2 literals). Here, the description consists of 2 properties $\{p, q\}$, such that $disc(p) + disc(q) = |C|$ or equivalently, $rulesOut(p) \cup rulesOut(q) = C$. However, this only means that the discriminatory power of the conjunction of p and q is maximal (i.e. rules out all distractors). Since GR searches among literals, it is possible that a property r exist such that $disc(r) > disc(p)$ or $disc(r) > disc(q)$. This would mean that r is selected by GR first. However, since a description must subsume the minimal description to be distinguishing, GR would keep up the search to yield a final outcome (at least as lengthy as) $\{p, q, r\}$.

Note that these two algorithms do not distinguish between coextensive alternatives of equal length. This is a consequence of the adequacy criterion in (2.5), which is purely quantitative. This is only problematic if there are properties in the KB which are inherently better than others, in the sense that speakers would be more likely to select them, or readers would find a description containing these properties easier to comprehend and resolve. It turns out that this is indeed the case; moreover, speakers are not 'Gricean', or not in the strict sense in which Dale (1989) interpreted the Maxim of Quantity. This observation was a core motivating factor for the Incremental Algorithm, proposed as a better alternative to the brevity-oriented strategies discussed here. Before turning to this algorithm and its many descendants, I take a detour through the psycholinguistic evidence that in part gave rise to it.

[5] Thanks to Chris Mellish for pointing this out.

2.6 Beyond functional relevance

Taken collectively, the psycholinguistic literature on reference production constitutes a falsification of the model based exclusively on functional relevance. By the mid-1980s, there was a significant body of such work, partially as a result of the establishment of an experimental referential communication paradigm by Krauss and Weinheimer (1964, 1966, 1967). In this paradigm – still dominant in current research – participants are presented with a 'visual world' consisting of a domain of abstract shapes or familiar objects, and participate in a game which requires them to refer to elements of this domain.

2.6.1 Are speakers Gricean?

Early developmental studies on reference sought to elaborate the observations made by Olson (1970) on the functional relevance of descriptions. Olson had proposed that, given the contrastive nature of referential communication, an increasing tendency to produce informative but non-redundant references should be evinced as a function of increasing age, because mature speakers have a better command of the communicative rules involved. The results of several studies over the next three decades suggested otherwise: while children's ability to refer successfully (i.e. produce identifying or distinguishing descriptions) improves with age, the tendency to produce overspecified expressions also increases, implying that functional relevance is tempered by cognitive constraints.

One of the earliest observations was that sensitivity to context, which enables the selection of distinguishing attributes for an intended referent, varied systematically with age. Adult listeners tended to be aware of referential ambiguity of descriptions, either requesting clarification when possible (Krauss and Weinheimer, 1966), or identifying a referent on a probabilistic basis when the task did not enable this (Rosenberg and Markham, 1971). Young speakers' ostensive lack of awareness of ambiguity suggested that they lacked the cognitive resources to carry out the exhaustive comparison (assumed to be necessary by these authors) between target and distractors to find distinguishing attributes. For instance, Glucksberg et al. (1975) cite an unpublished study by Glucksberg and Kim, in which children were asked to instruct a confederate to stack coloured blocks on a peg. The children were as likely to use functionally irrelevant attributes as they were to use relevant ones, suggesting that they were not carrying out the comparison process. Similarly, Ford and Olson (1975) tested young children against an older control group in a task resembling the Krauss/Weinheimer dialogue game, in which the complexity of the description required to identify an object (in terms of number of properties) was systematically varied. Younger children tended to produce more underspecified descriptions than older ones. Interestingly, Ford and Olson also found a tendency for older children to produce longer descriptions than necessary. However, they argued that such descriptions, although overspecified in the context of a specific domain, were adequate in the context of the experiment as a whole, in the sense that the redundant attributes used in a given trial were contrastive in relation to distractors on previous trials. This argument left open the possibility that as children got older, they tended to observe the requirements of the Gricean Quantity Maxim.

A later study by Whitehurst (1976) compared referential communication of children in four age groups. In his Experiment 1, children were shown domains consisting of cups varying in size, colour or pattern, one of which was marked as an intended referent. The size of the domain (i.e.

the number of distractors – 1 or 2) and the length of the minimal description required to identify the referent were systematically varied. Whitehurst reported an increase in the probability of successful references with age, with a concomitant increase in overspecification. In a second experiment, a group of first-grade children were trained in the reference task by an adult model. Although the success rate increased as a result, it did not reach significance, forcing the conclusion that children operate on a 'principle of least effort'. Similar conclusions were reached by Whitehurst and Sonnenschein (1978). Both of these studies also found an effect of the complexity of the task, that is, the cost (gauged by the likelihood of an unsuccessful reference) incurred by requiring more attributes to distinguish a referent, or having a greater number of alternatives to choose from. In a later study, in which children of different age groups were required to identify a stimulus array based on a description of its contents, Sonnenschein (1982) also reported an effect of stimulus complexity. A further manipulation involved the degree of overspecification of messages that children were given in a reference resolution task. Sonnenschein makes the striking conclusion that overspecification *hinders* younger children, while older ones *benefit* from it, especially in case the stimulus array was very complex.

In these early studies, the tendency to overspecify was explained either in terms of a procedural deficiency (children not having mastered the contrastive function of reference), and/or in terms of a principle of 'least effort', as shown in the following quotation:

> it is far easier to be redundant than efficient [...] There is little reason to expect minimal redundancy to be a routine attribute of communication *at any level of development*. Unless there are specific reasons to behave differently, children seem to operate on the principle that words are cheap. (Whitehurst, 1976, p. 482; emphasis added)

A further factor found to improve communicative efficiency was the presence of a communicative partner that gave feedback, and the possibility of alternating the subject's role from that of speaker to that of listener in the course of the communicative task, as it were giving them 'direct experience' of the listener's role (Krauss and Weinheimer, 1966; Deutsch and Pechmann, 1982; Sonnenschein and Whitehurst, 1984; Sonnenschein, 1984). Although the latter finding does not constitute direct evidence for speaker-listener asymmetry, it does suggest that young speakers may be unaware of what (in the Gricean model) is helpful to listeners. More direct evidence for such an asymmetry is found in a recent study by Engelhardt et al. (2006), reviewed below.

While the influence of a communicative partner is highly plausible (cf. Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996), these early studies tended to manipulate such communicative situations in order to provide a social model for the child. The motivation was a rather normative view of what constitutes correct communication, which raised the question of whether an adult model might be able to influence the child's behaviour. Moreover, an explanation of overspecification in terms of procedural deficiency fails to explain why the tendency increases with age, while appealing to a principle of least effort begs the question as to the cognitive procedures that incorporate such a principle. Another possible criticism of these studies is that the experimental procedure used often drew the attention of young children to one specific referent in a domain or visual world, possibly causing them to pay less attention to distractors in context (cf. Lloyd and Banham, 1997).

Later work sought a more mechanistic explanation of overspecification, by focusing on the incremental nature of language processing. Recall from §2.2 (p. 25) that Wundt's Principle predicts that a module in the language processing pipeline will respond to a minimal amount of the correct input. This would imply that in the standard referential communication paradigm (visual world + intended referent), a minimal amount of perceptual input will initiate an utterance, meaning that speakers do not compute the 'discriminatory power' or contrastive relevance of a property before uttering it. It would also suggest that if some properties of a referent are more perceptually salient than others, they will tend to feature in an utterance irrespective of whether they have contrastive relevance, because they are perceived earlier. In terms of the model of Chapter 1 (Figure 1.2, p. 18), perceptual salience increases the likelihood that a property form part of the mental representation of a referent. For the same reason, gradable properties, such as SIZE, would evince the opposite tendency, because in order to use a property such as ⟨SIZE : *large*⟩, it is necessary to make an explicit comparison of an object to the surrounding context. Another prediction that falls out of the incremental model of language production is that certain properties – specifically, TYPE[6] – have a privileged status, not only because of their perceptual salience and role in object recognition, but because they are important building blocks for the incremental NP-construction process, assuming this process to be head-driven (cf. Kempen and Hoenkamp, 1987).

These hypotheses have received widespread support, starting with a series of studies by Pechmann (Pechmann, 1983, 1984; Schriefers and Pechmann, 1988; Pechmann, 1989). In an early eye-tracking experiment, Pechmann (1984) showed that speakers begin the articulation of a message before their scanning of the visual domain is complete. The speech data also showed that the syntactic constraint that SIZE be expressed before COLOUR (in Dutch, as in English) was violated quite often, suggesting that people were expressing the COLOUR property before the gradable property. Later work showed that TYPE and COLOUR were always used in referring expressions, even when they had no contrastive value (Schriefers and Pechmann, 1988; Pechmann, 1989). This result was not replicated for SIZE. The privileged status of TYPE was argued by these authors to be due not only to syntactic processes, but also to the fact that speakers process a referent as a **conceptual gestalt**, central to which is the referent's object class; similarly, they argued that COLOUR had both perceptual and conceptual primacy, forming part of the gestalt that is the speaker's mental representation of an object. Similar results have been reported by Mangold and Pobel (1988), and Eikmeyer and Ahlsèn (1996). In the latter study, on German and Swedish speakers, the proportion of minimally specified references was extremely low (ca. 3.5%), while COLOUR was used in the vast majority of cases, irrespective of its contrastive value.

Why are some properties preferred over others? The explanation suggested by the above studies is that properties of a referent which do not require explicit comparison with other objects are faster to process and feature early in incremental production. Moreover, some properties are more 'intimately' bound to the conceptual representation of an object: COLOUR, for example, is an inherent property of an object, while SIZE is relative. Some evidence for the primacy of COLOUR in object representation has been reported by Naor-Raz et al. (2003), who found evidence that colour perception facilitates object recognition. More directly relevant to the present discussion is a study

[6] Alternative names for this attribute include *object class* and *conceptual category*.

by Belke and Meyer (2002), which directly addresses the question of perceptual/conceptual primacy of attributes in reference. Belke and Meyer coupled an eye-tracking methodology with a *same-different judgement* paradigm, in which participants are shown a target and a context object and respond by indicating whether the target is different from the context object, or whether they are the same. A robust finding in this paradigm is that *different* judgements tend to be significantly faster, indicating that search is self-terminating, with responses made as soon as sufficient information has been scanned. In the *same* condition, this requires an exhaustive comparison, since every attribute of the objects has to be compared. However, this tendency depends on the nature of the attributes involved. The advantage of *different* conditions disappears if the distinguishing features of an object have a low degree of **codability**, that is, require comparison to other objects to determine their value. Belke and Meyer hypothesised that codability should also interact with *discriminability* of an attribute. Thus, an attribute like SIZE has relatively low codability, but a large difference in size between two objects would make the difference salient, and hence reduce the latency incurred by the codability effect.

Belke and Meyer's main experiment involved objects defined in terms of COLOUR, TYPE and SIZE. They manipulated the number of attributes along which two objects differed in the *different* condition. In case SIZE was contrastive, the size difference ratio was $5 : 4$ (i.e. hardly codable). The results indicated more complex viewing patterns (gaze shifts between the two objects), and longer response latencies, in the *same* condition overall. However, SIZE differences gave rise to much longer latencies, and more complex viewing patterns in the *different* condition. Crucially, the effect of SIZE only appeared when it was the only contrastive attribute, that is, when TYPE and/or COLOUR were contrastive in addition to SIZE, the latter simply did not enter into the referential equation. This not only supports the self-terminating search hypothesis, but also goes some way towards explaining (in terms of codability) Whitehurst's (1976) observation of a principle of least effort. The authors extended these findings to a referential communication task, proposing to view it as a series of *same-different* judgements, whereby speakers determine whether a property has contrastive value or not. They suggested that the following three-stage model would be required to produce a minimal description:

1. Detecting differences between target and distractors;

2. Evaluating differences in terms of discriminatory power;[7]

3. Verbalising the distinctive features.

Note that this model precisely echoes the FB strategy, and involves exhaustive comparison, which Belke and Meyer suggested was humanly intractable (as it is computationally). Their hypothesis was that low-codability properties like SIZE would in fact be filtered out at Stage 1 wherever possible, leaving only high-codability features for later evaluation. This hypothesis was supported: A high proportion of overspecified references was observed, with more use of COLOUR overall, even when it was not contrastive (80%). However, speakers seldom used SIZE in overspecified references. Taken together with Belke and Meyer's online visual processing experiment, this study constitutes a very direct falsification of the FB model of reference.

[7] The term used by the authors is *distinctiveness*. I have used *discriminatory power* to maintain consistency.

Further evidence of speakers' failure to directly evaluate discriminatory power comes from Pechmann (1983, 1989), who analysed the accentuation strategies of speakers in his experiments, based on the hypothesis that information marked as contrastive would be stressed. He found that both children and adults tended *not* to accentuate properties of an object that distinguished it from its immediate distractors (so-called **exophoric contrast**). However, when the experimental paradigm involved multiple references to objects within the same domain, clear evidence was found for **endophoric** contrast, that is, speakers accentuated the properties that distinguished the current object from previously mentioned referents. This lends further support to the incremental hypothesis. When speakers verbalise a property, they are not yet aware of its contrastive value in the immediate context; however, memory for the preceding discourse would cause them to contrast a property with previous referents (cf. Levelt, 1989, chap. 4).

2.6.2 Perception-production coupling in referential communication

As noted above, the incremental model of language production is based on the premise that the perception, conceptualisation and NL production processes are closely time-locked, with perceptual input feeding into language production modules at the earliest possible stage. Direct evidence for this comes from a series of eye-tracking experiments in the Visual World Paradigm by Tanenhaus et al. (Tanenhaus et al., 1995; Eberhard et al., 1995; Chambers et al., 2002), essentially a version of the Krauss/Weinheimer referential communication task, in which participants' gaze shift is tracked. The basic idea behind the eye-tracking methodology is that saccadic eye movements signal shifts of attention. Since it is well-known that initiation of a saccade takes approximately $200ms$, these experiments permit a very accurate picture of the time-course of online processing and domain circumscription in reference resolution. The task given to subjects usually involves resolving a reference in order to carry out an instruction, and the focus is on the way attention shift changes as a function of language comprehension.

The Visual World paradigm has usually focused on listeners, and has consistently shown that shifts of attention are highly dependent on the incoming linguistic signal. One important finding is the *Point of Disambiguation* (POD) effect, whereby listeners' gaze stabilises on a target object as soon as a property is mentioned that disambiguates it from its distractors. Thus, in a domain with two objects, both apples, only one of which is red, listeners home in on the target referent as soon as the word *red* in *pick up the red apple* is uttered. The POD has also been shown to alter the normal effect found in visual search paradigms, which require subjects to search a visual domain for an object with specific properties. In the purely visual paradigm, where subjects are given the instructions before search, the time taken to find an object varies systematically with the number of distractors in the domain. When the task is accompanied by instructions *during* search, the effect of number of distractors is reduced (Spivey et al., 2001), implying that people circumscribe the referential domain incrementally. Further results by Sedivy et al. (1999) show that POD generalises to certain vague or gradable predicates. Apart from their low-codability status in Belke and Meyer's sense, these predicates are known to manifest 'global' dependencies. Thus, a person might be referred to as *the tall woman* if she was tall for a woman, implying that there is a standard of comparison for the applicability of the word *tall* to an object of a given class. On the other hand, Sedivy et al. found no significant difference in response latency or gaze shift, as a function of whether an adjective applied intrinsically to an object or not (e.g. whether a target referred to as *the tall cup* was tall for a cup). This finding implies that even the processing of gradable properties is dependent in large measure on the immediate perceptual context.

The POD effect actually raises a challenge to earlier results on overspecification. Let us assume that both incremental production and resolution involve self-terminating search. The difference is that the listener is in a position where her search is guided by a linguistic signal, whereas the speaker's utterance is only guided by a post hoc evaluation of whether the reference produced at a given stage is sufficient for identification, *modulo* the preference for visually salient properties and constraints on conceptual representation (Pechmann's *gestalts*). If this is the case, then it is possible that speakers' utterances will be overspecified, but that this may hinder listeners' resolution processes, since POD implies that their search will terminate on encountering a disambiguating property.

A recent study by Engelhardt et al. (2006) in the Visual World Paradigm presented subjects with instructions to move a target object that was on or in another object to a new location. For example, the target might be an apple placed on a towel. Thus, instructions contained two references, one to the target (e.g. *the apple on the towel*) and the other to the target location (e.g. put the apple on the towel *in the pot*). Two variables were manipulated: (1) the new target location object was of the same type as the target's initial location (*matching* condition) or not (*different*); (2) a distractor of the same type as the object was present or absent. The main focus of the studies was on whether speakers and listeners would evince a preference for overspecified NPs. In the case of the target reference, overspecification would involve using a PP modifier (*on the towel*) when no distractor was present. Overspecification of the target location reference would involve using an operator like *other* (*put the apple on the towel on the other towel*).

Two initial offline experiments involving speakers and listeners respectively showed a preference for overspecified descriptions. Speakers produced overspecified instructions to a confederate 30% of the time in the absence of a distractor. Whether the target location was *matching* or *different* had no effect. Listeners carried out a meta-linguistic task, judging the adequacy of an instruction to carry out the requisite action. Overspecified references were rated as no worse than non-overspecified ones, whether or not there was a distractor. However, subjects did penalise underspecified target location references in the *matching* condition. Thus, an expression like *put the apple on the towel on the towel* was rated negatively, compared to *put the apple on the towel on the* other *towel*. Incidentally, this example illustrates a potential confounding variable in the experiment, since the use of *other* might be required on independent grounds in a context like this one, perhaps because semantically it is an alternative-set operator, which explicitly establishes a link to a previously introduced entity of the same object class (Bierner and Webber, 2000).

Engelhardt et al.'s online follow-up experiment replicated the POD effect, with significantly more looks to the target once the head noun was uttered. Gaze shifts to the target location were somewhat different. When no prepositional phrase modifier was used with the target reference, shifts to the target location were slow, and slowest in the *matching* condition (e.g. *put the apple on the towel* in a domain where the apple is on a towel and there is another towel). Overspecified expressions on the other hand caused some confusion, with subjects shifting between target location and original location in the *matching* condition. This apparent cost contrasted with the results of the earlier two experiments, in which speakers did produce overspecified expressions in the *matching* condition, while listeners judged them as being no worse than underspecified ones. The conclusion reached by the authors is that producing non-overspecified descriptions is computationally costly for a speaker, whose mental representation of the target referent includes its location (thus, *the apple on the towel* rather than just the *apple*). This is essentially an extension of Pechmann's concept of a *gestalt* to include locative properties. However, the authors also note a speaker-listener asymmetry, with overspecification (at least of the sort tested here) being costly for a listener to process, in spite of listeners' metalinguistic judgements to the contrary.

In a similar vein, Arts (2004) found that while location information facilitated reference resolution for her subjects, speakers in the same situation did not tend to produce locative expressions to the same extent, possibly because relational attributes such as location are costly for a speaker (they involve relations between multiple objects). This amounts to the same explanation as Engelhardt et al.'s for the opposite effect: here, overspecification is useful for the listener, but costly for the speaker. The benefits of overspecification for listeners are also reported in a recent study by Paraboni et al. (2006), focusing on hierarchical domains such as documents.[8] Participants were presented with an electronic document, and were asked to resolve references to different document parts. Effort was measured as the number of clicks a subject made to reach the right part of the document, and a clear benefit of overspecified reference was found.

In summary, there is a robust tendency on the part of speakers to overspecify. However, it may depend on the kind of attributes involved, and its benefit for listeners is contingent on the type of domain (Arts, 2004; Paraboni et al., 2006), and the nature of the distractors (Engelhardt et al., 2006). Somewhat less clear-cut is the possibility that what is easy for speakers is not always beneficial to listeners.

2.6.3 Speaker-listener asymmetries and communicative intention

Evidence for speaker-listener differences is problematic for GRE algorithms that are reliant on theories of communicative action based on a principle of cooperation (Grice, 1975) or on mutual recognition of communicative intentions (Searle, 1969), since the assumption of symmetry is central to these theories. This point has been raised in GRE by Oberlander (1998), whose review of some psycholinguistic evidence leads him to conclude that endorsing some version of the Gricean maxims in GRE might be a risky strategy, because the principle of 'doing the right thing' is measured by different yardsticks depending on what communicative role one takes. We have already seen some evidence that certain linguistic mechanisms (overspecification, relations), are beneficial to different degrees for speakers and listeners. What is the evidence that referential communication is based purely on a collaborative effort to understand an interlocutors' intentions and establish common ground?

The theory that interlocutors' referential acts are constrained by considerations of mutual knowledge has been prevalent at least since Clark and Wilkes-Gibbs (1986). Some authors, notably Keysar (1997) have argued that early experiments purporting to demonstrate this were confounded because the presence or absence of shared knowledge, and mutual awareness thereof, was not directly manipulated as an experimental variable. Recent work by Keysar et al. (2000, 2003) has yielded some preliminary evidence that interlocutors may have difficulty in taking full account

[8] Such domains are hierarchical in the sense that they can be represented as a tree, with some parts of the document subsuming others.

of which knowledge they know to be shared, often prioritising their own 'privileged' ground in resolving a partner's utterance. Nevertheless, the evidence here is still tentative, with other studies showing that conversational partners do align in their referential communication, in various ways. Hanna et al. (2003) and Hanna and Tanenhaus (2004) report evidence from the Visual World eye-tracking paradigm that subjects do focus on what they are aware is common ground, paying less attention to possible referential targets which they know are not mutually available. Brennan and Clark (1996) found that interlocutors begin to align their use of lexical items, using the same words to refer to objects. This kind of **lexical entrainment** has been cited as one way in which interlocutors align their knowledge states (Pickering and Garrod, 2004). Metzing and Brennan (2003) report evidence that interlocutors develop implicit conceptual pacts with specific partners, and are are misled when partners break these conceptual pacts, suggesting that interlocutors tend to be cooperative and also expect their partners to observe the cooperative rules established temporarily during a conversation.

This body of work can be interpreted in two ways. On the one hand, researchers in the *language-as-action* paradigm would view it as evidence for a conscious, collaborative effort on the part of interlocutors to achieve alignment; conversely, the evidence might be interpreted in terms of a more primitive, unconscious priming mechanism operating at different levels, from the linguistic to the conceptual (Pickering and Garrod, 2004). Whichever interpretation one selects, the evidence must at this stage be considered inconclusive.

Some of the experimental results reviewed above could be strongly dependent on the communicative task. For instance, casual dialogue might permit interlocutors to be less rigorous in taking listeners' perspectives into account. Indeed, extensions of the Visual World paradigm to naturalistic dialogue (Brown-Schmidt et al., 2002; Campana et al., 2002) revealed that a significant proportion of NPs apparently were *under*specified, but that this did not hinder participants from focusing attention on the right target. It turned out, however, that the apparent underspecification was due to the referential context having been constrained during the previous discourse, so that only a subset of the objects in the domain were in focus. Another factor might be that the situations tested were not fault-critical. von Stutterheim et al. (1993) report an experiment which required participants to describe an unfamiliar object, either in a scenario where they had to *instruct* someone to reconstruct the object from its parts, or in a scenario where they simply had to describe it. In their analysis, the authors focused on the use of COLOUR as a property in NPs which *introduced* a novel object, *maintained* an object in focus from a previous utterance, or *re-mentioned* the object later. In the *description* condition, COLOUR was equally distributed throughout these classes of NPs; however, the *instruction* condition showed a marked decrease in overspecification in *maintenance* NPs, compared to NPs which introduced or re-mentioned a referent. The difference between *description* and *instruction* conditions was explained in terms of how the objects are conceptualised. An instruction aims for an interlocutor to identify a referent in order to manipulate it, and is thus more fault-critical, giving rise to greater overspecification in initial references and re-mentions, while reducing the necessity to overspecify once the object is in focus. Similarly, Arts (2004, chap 5) found that participants who had to describe a radio panel to an imaginary interlocutor, produced twice as many overspecified references to the buttons on the panel when the goal was instructive. These references were exhaustive (using all three possible attributes of SHAPE, COLOUR and SIZE) 33% of the time.

Thus, the communicative task does seem to affect the way speakers formulate references. Further evidence comes from corpus-based studies. van Vliet (2002) claimed that conceptual shifts in narrative texts (e.g. a shift in perspective on the referent, or a turn of events) were marked by overspecified NPs. If this is correct, then it is evidence for **intentional overload** (Pollack, 1991), whereby the 'prototypical' form of an utterance with a specific communicative intention (definite NPs used for reference) is used to satisfy a further intention apart from the basic one. Jordan (2000b,a, 2002) made a number of studies on the COCONUT dialogue corpus, and found that interlocutors' economy in referring to objects in a shared domain of conversation varies depending on the task.[9] For instance, a person might repeat a partner's reference to an object in order to signal agreement with their interlocutor to buy that object, something that a model based purely on functional relevance would not predict. Similarly, if a speaker wanted to adjust task constraints, they might opt to use an overspecified expression containing properties that might allow the hearer to infer the proposed adjustment. In Jordan's *Intentional Influences* model, these results are interpreted as evidence for communicative intentions influencing the content of referring expressions. This seems plausible, given the well-defined nature of the task in COCONUT. However, the results also leave open the possibility that speakers are simply doing what is easiest for them. For instance, repeating a partner's reference could be the least effortful way of referring to the object currently in focus. Moreover, speakers' intentions in these studies were imputed to them in a *post hoc* fashion, rather than directly manipulated.[10] Although the inference that a speaker had certain intentions at a given point in a dialogue is plausible, given the well-defined nature of the task, there is still the possibility that the effects observed are not caused by intentional shifts. Nevertheless, a machine learning study by Jordan and Walker (2000, 2005) did find that the Intentional Influences model made the correct predictions about the attributes to be selected for NPs, compared to other computational models (see Chapter 4).

2.6.4 Summary: Was Grice wrong?

The past few decades have witnessed a vast amount of psycholinguistic research on referential communication, both on speakers and listeners. Some of the findings in this literature are strikingly robust; others less so. The preceding discussion can be summarised as follows:

- Reference production and resolution are incremental, self-terminating processes. Rather than exhaustively analysing a domain for a distinguishing description, speakers initiate references as the search unfolds.

- Incrementality also challenges purely intentional models such as one based on Grice (1975). The evidence is that while speakers may be driven by their communicative intentions, and influenced by the type of domain at hand, what they achieve depends in part on automatic processes involved in production/comprehension.

- There are potential asymmetries between speakers and listeners in terms of what facilitates production/resolution, which may be due to the different mechanisms at play when a person

[9] COCONUT dialogues are task-oriented, and were collected using a game with well-defined rules and constraints.

[10] The *post hoc* logic is arguably characteristic of any corpus-based study, unless the corpus in question is experimentally designed to falsify some hypothesis. See Chapter 3 for further discussion.

occupies one or the other role.

The psycholinguistic data raises the obvious question of whether the Gricean maxims have any validity, at least in the context of reference. This depends on how Grice is interpreted. Dale and Reiter (1996) suggest that the maxims should be taken as *post hoc* observations of communicative tendencies. Similarly, Bach (2005) lists the tendency to regard the maxims as directives or empirical predictions among his 'top ten misconceptions on implicature'. Although Bach's point is well taken (that is, Grice never intended his maxims as 'directions'), there is another dimension to this debate related to the possible tension between principles stated at an *intentional* level, and observations of how the language production mechanism operates. While Grice's maxims may reflect people's communicative intentions to some extent, the way these intentions are ultimately realised will also depend on the available machinery. To draw an analogy, one may have the intention to get from one place to another in the shortest possible time, but the realisation of that intention depends on what means of transport ('machinery') are at one's disposal. Similarly, even if speakers and/or listeners may judge brevity to be a virtue, other mechanisms (of the sort that Pechmann and others have described) may play a strong determining role in what speakers actually produce (as witness the results of Engelhardt et al. (2006)).

Perhaps a better way to view the apparent clash between maxims stated at the pragmatic/intentional level, and actual behavioural tendencies, is to factor in the cognitive mechanisms underlying the latter. If automatic processing is in part determined by built-in mechanisms, then deviations would be expected from what are ultimately observations stated at a different level of explanation. This argument amounts to a case for taking the insights of both the language-as-action view of Clark (1996) and the 'opposing' language-as-product view into account. Similar issues will arise later in this thesis, where the discussion of the Conceptual Coherence model (see especially §7.6 and §7.7) centres on 'low-level' lexical issues, but raises questions regarding the interaction of these with speakers' overall communicative goals and the perspective they intentionally take on elements of their domain of discourse.

Against this psycholinguistic backdrop, let us now shift our attention back to the computational literature, and to how GRE began to view the role of the Gricean maxims as a softer constraint than hitherto assumed.

2.7 The Incremental Algorithm and its descendants

The Incremental Algorithm (IA) by Dale and Reiter (1992, 1995) had two motivations: (a) the computational intractability of a strict adherence to the Gricean Maxim of Quantity; (b) the observation that such an adherence is in fact relative when human speakers are observed in real or laboratory situations. IA was proposed as a superior model of reference, compared to FB, GR, and LB, both for reasons of efficiency, and for its better reflection of the psycholinguistic data.

There are two senses in which IA is incremental. First, like GR, it adds properties to a description as search unfolds, rather than conducting exhaustive search. This means that, as in GR, search is self-terminating; the description is returned as soon as it has excluded all distractors in the domain. The second sense in which it is incremental (and where it diverges significantly from GR), is in the order in which properties are searched. Rather than calculating discriminatory power at every stage, IA's search procedure is based on a *a priori* order of properties, sometimes called the **preference order** (PO), in which highly preferred attributes are placed earlier than others. This makes IA not a greedy, but a **gradient descent** algorithm.[11] The order in which properties in the search space are tested is determined by the attribute whose value a property represents, relative to the PO, which is represented as an ordered list. Let $index(\text{A}, \text{PO})$ be the position of attribute A in this list. Then the ordering relation among properties that characterises the IA is defined as follows:

$$\langle \text{A} : v \rangle >>_{p_{\text{IA}}} \langle \text{A'} : v\text{'} \rangle \leftrightarrow index(\text{A}, \text{PO}) < index(\text{A}', \text{PO}) \tag{2.13}$$

that is, $\langle \text{A} : v \rangle$ is preferred to $\langle \text{A'} : v\text{'} \rangle$ if A is ordered in PO before A′. In terms of the heuristic search framework introduced in §2.4 (p. 31), the return value of the $dequeue(Q)$ function of the IA is defined as follows:

$$dequeue(Q) =_{def} \arg\min_{\langle \text{A}:v \rangle \in Q} index(A, \text{PO}) \tag{2.14}$$

If it is assumed that the queue contains all properties in P_r, ordered in this manner, then the procedure of the IA is identical to that shown for GR in Algorithm 2, the only difference being the ordering relation. Thus, the procedure starts by initialising the description D to the empty set, and then proceeds along the PO of attributes. At every stage, a property is tested for whether it excludes some distractors, in which case it is added to the description D, and the context set C is updated (see Algorithm 2, p. 36). As an example, consider a reference to e_7 in Table 2.1 (p. 28). Assuming the preference order shown in the top row of the table, the algorithm would terminate immediately after considering TYPE, since *dog* is distinguishing. On the other hand, under this PO, a reference to e_2 would contain some overspecification, as witness (2.15), compared to the minimally specified (2.16).

(2.15) $\langle$TYPE : *woman*$\rangle \wedge \langle$ROLE : *maid*$\rangle \wedge \langle$CLOTHING : *wears_grey*$\rangle$

(2.16) $\langle$POSTURE : *kneeling*$\rangle$

Another feature of the IA is that it explicitly includes a function to deal with a further aspect of the pragmatics of reference, namely, the preference for basic-level categories in describing objects (Rosch, 1973). This preference, whereby people prefer terms such as *dog* to more specific terms such as *terrier*, was argued by Cruse (1977) to reflect a Gricean tendency to avoid false implicatures, since more specific terms, used when a basic-level term would suffice, give more information than required. Under the assumption that some ontological or taxonomic support structure is available in addition to the KB, Dale and Reiter proposed a function *findBestValue*(A), which attempts to find, for a given attribute A, the value which is closest to the basic-level, and which is also discriminatory (i.e. removes some distractors).

As a consequence of the predetermined order of search, the resulting description can be overspecified. This, however, is justified on the grounds that overspecification is what people do anyway; moreover, as we have seen, speakers manifest preferences for some attributes over others. Dale and Reiter argued that this made IA more psycholinguistically plausible. Additionally, the

[11] An alternative name for this procedure is **hillclimbing**.

PO allows the system designer to incorporate domain-specific preferences in the algorithm, so that attributes which are very relevant to a particular domain of discourse for which an NLG system is designed will be more likely to be included in descriptions. The notion of 'preference' has also been interpreted as reflecting the 'cost' involved in describing a referent by including a particular attribute, and has proved useful in extensions of GRE to new scenarios, such as multimodal reference, where the decision to include a property has to be balanced against other decisions, such as whether to also accompany a definite description by a pointing gesture. Krahmer and van der Sluis (2003) propose an algorithm that achieves this by estimating how costly or effortful it is to add a property to a description which is not distinguishing, compared to accompanying the description by a pointing gesture. The basic idea is that, like attribute-value pairs, pointing gestures can also be ordered in terms of cost[12], thereby allowing the estimation of overall effort involved in using properties or gestures to proceed in a unified fashion.

The argument that the IA is psycholinguistically plausible is rather tenuous. Evidence for speaker preferences for certain attributes is based on the observation that such attributes are included *whether or not* they have any contrastive value for the referent. Thus, an algorithm could only be said to respect people's preferences if every description that the algorithm generates contained the preferred attributes (assuming that such attributes are known in advance, an assumption that the authors make). Whether the IA actually does this crucially depends on the PO. To take an example, suppose that COLOUR and TYPE are preferred attributes, and are therefore placed first in the PO, in the order TYPE $>>$ COLOUR. This would only guarantee that they are *considered* before other attributes. If TYPE alone were sufficient to distinguish the referent, then the algorithm would never consider COLOUR as a potential attribute for inclusion. Mindful of this possibility, the authors propose a further function that checks whether a description returned by IA actually contains a TYPE attribute. If not, then one is added at the end of the procedure, because, as discussed in §2.6.1, this is the core part of the representation of an object. However, whether or not the overall procedure in this example approximates the psycholinguistic data is questionable, if it turns out (as the data indeed suggests) that COLOUR would be included anyway by a speaker in the same situation. This problem is an instance of a more general feature of gradient descent algorithms, namely their susceptibility to local maxima. In the current example, TYPE is such a local maximum, since on reaching this point, the algorithm terminates with success, omitting COLOUR which (perhaps) should also have been included. Perhaps this problem arises as a result of combining different adequacy criteria, namely, the extensional success criterion in Definition 3 (p. 31) on the one hand, and the desirability of including preferred properties. Some authors, notably van der Sluis and Krahmer (2005) and Horacek (2005), have proposed frameworks in which properties which have no contrastive value are added to a description if they have very low cost (van der Sluis and Krahmer) or reduce the uncertainty associated with the intended referent of a description (Horacek). However, these proposals have tended to be of a theoretical nature, and have not benefited from extensive empirical evaluation.

It is worth noting that the reliance on a predefined PO means that the behaviour of the algorithm varies, not only as a function of the KB, as one would expect, but also as a function of how properties are ordered, giving the algorithm a certain degree of non-determinism. It could in

[12] This is based on factors such as proximity to a target referent, and how they affect the precision of the gesture.

fact be argued that in a KB with n different attributes, there are $n!$ possible orderings and therefore $n!$ possible IAs. This makes a formalisation of the descriptive adequacy criterion $>>_{D_{IA}}$ less straightforward than it was for previous algorithms, in part because, while the ordering relation (2.13) is well-defined for any $\langle \text{A} : v \rangle$ in the KB, the description itself may or may not contain highly-ranked attributes for the reasons just outlined. Unlike FB and GR, whose descriptive adequacy criterion is theoretically-motivated, and makes a judgement of their success relatively straightforward, the same criterion in the IA is completely domain-dependent.

The second case for the IA made by Dale and Reiter is based on its efficiency, and rests on more solid formal grounds. The IA is arguably more efficient than GR. The latter is polynomial in $|P_r|$, the number of properties true of the intended referent. IA does not perform online comparisons between properties, because of the PO. If we assume that there are at most n_a attributes in the KB, each with at most n_v values, then IA has complexity $O(n_a n_v)$. If, on the other hand, we assume that the PO consists of the attribute-value pairs in $P_r \subseteq \mathbb{P}$, ordered as per (2.13), then IA is linear in $|\mathbb{P}|$.

To summarise, IA was based on a relative interpretation of the Maxim of Quantity, and thus contrasts with earlier interpretations. Its attractiveness lies in that it is highly efficient and, at least in some cases, produces overspecified expressions that conform to speaker behaviour, although the latter feature cannot be guaranteed. Further developments in GRE since the publication of this algorithm have often taken it as their starting point. In what follows, I focus on three lines of such research. One has focused on the refinement of the notion of the contrast or distractor set, making the generation process sensitive to discourse context. Contrast sets did not take centre-stage in Dale and Reiter's exposition, but the form and informativeness of referring expressions is known to be highly dependent on whether their referents are being attended to by a speaker. The second set of developments has to do with **expressiveness**. Proposals have aimed to extend the remit of GRE algorithms to deal with different kinds of predicates. In §2.7.3 and §2.7.4, I discuss two such extensions, to n-ary relations and numeric-valued attributes. The third line of research has focused on **logical completeness**, extending coverage to deal with negation and references to multiple entities, with particular focus on plurals. This is particularly pertinent to the present work, which focuses on plurality.

Although many of these extensions took the IA as a starting point, they could just as easily be accommodated by the other algorithms, given a common formal framework such as the one used here.

2.7.1 Context

The dominant focus in research on context in GRE is on information reduction in referring expressions. If a referent has been introduced earlier in a discourse, then subsequent references to it should reflect this, especially if the referent was mentioned quite recently.

One of the earliest proposals for incorporating some context-sensitivity in GRE was by Dale (1989), whose algorithm used the simple heuristic of referring using a pronoun if a referent had been introduced in the previous sentence. Context also played some role in the IA, insofar as Dale and Reiter suggested that the distractors from which an entity is distinguished are the most salient (i.e. in the focus of attention) in the discourse context. This was however one of of the less elaborated aspects of the IA in its original formulation. The influence of discourse on NPs is an area of intensive research. Here, I will focus on those aspects that are directly relevant to GRE.

Central to most theories of anaphora is the notion of **salience**. A related notion, introduced by Ariel (1988), is that of **accessibility**, but it is probably more correct to say that salience is one of the determinants of the accessibility of a discourse entity (cf. Ariel, 2001, for a discussion of other factors influencing accessibility). Salient entities warrant a reduction in the amount of information used to refer to them; the usual candidates are reduced definites and pronouns. Determinants of salience include recency of mention, and the discourse function of an NP. In Centering theory (Grosz et al., 1995), salience is determined mainly on the basis of the grammatical role of an NP denoting a discourse entity, with the following hierarchy of salience: SUBJECT >> OBJECT >> OTHER. An alternative model (Hajiĉova and Vrbová, 1982; Hajiĉova et al., 1990; Hajiĉova and Sgall, 2001) views the salience of a discourse entity as a function of (a) its status as a focused or non-focused entity; (b) the distance from the last mention of that entity

There are two reasons why salience should play a role in GRE. First, the **local coherence** of a discourse is strongly influenced by the forms of the NPs used to maintain referents in focus (Grosz et al., 1995). Thus, the naturalness of output of an NLG system will depend in part on the context-sensitivity of its GRE component. Second, accounting for salience arguably involves extending the scope of GRE algorithms, to take into account issues related to realisation and possibly even text planning.

It is only relatively recently that generation algorithms have begun to take into account the finer-grained notions of salience in theories of discourse. Kibble (1999) proposed to incorporate aspects of Centering theory into NLG, via a division of labour between text planning and microplanning. While maintaining smooth transitions between discourse segments (global coherence) is the job of the text planner, the microplanning phase (and GRE in particular), was proposed as the locus for decisions on pronominalisation or NP reduction. This proposal introduces some elements of realisation in the GRE module, since the decision to pronominalise is one about form as well as content. It highlights a strong dependency between text planning and NP generation, one that is further elaborated by Kibble and Power (2000).

Arguments for a tighter coupling between text planning, realisation and GRE have also been put forward by Passonneau (1997), whose empirical work shows that there are correlations (albeit imperfect ones) between global discourse structure and NP realisation. Passonneau (1995) proposed an extension of the IA, which takes as the context set the entities in the current utterance, and attempts to refer to a target first by using a pronoun. If this fails (in case there are equally salient distractors with the same grammatical features as the target), the algorithm produces a description with only a TYPE, and adds modifiers only if required. The algorithm was incorporated into a model that also took global constraints into account. Similar arguments for making GRE more sensitive to discourse and realisation concerns are put forward by Dale (2003), for the specific case of *one*-anaphora. Dale argues that, apart from their anaphoric function, *one*-anaphors often serve to indicate contrast.[13] This is a rhetorical relation which falls within text planning; hence the decision to use such an anaphor could be taken as early as the text planning stage.

McCoy and Strube (1999a,b) also observed a dependency between salience and NP form.

[13] For instance, in *Dave bought a new Ford. Mary has an old one.*, it has been claimed that *one* contrasts Mary's Ford to Dave's on the basis of their age.

Their corpus revealed that the decision to pronominalise depended on the distance between anaphor and antecedent, as well as the *thread* of the discourse. Threads are sequences of utterances in which a specific discourse intention or theme (Grosz and Sidner, 1986) is evident. In Grosz and Sidner's framework, intentions played a causal role in structuring discourse into segments, and these authors proposed that full NP reference would be more likely at the beginning of a segment, which constitutes a focus space. However, Passonneau (1997) found this correlation to be poor in many cases; thus, McCoy and Strube proposed that rather than the stack of focus spaces in the original Grosz and Sidner theory, discourse should be characterised as having multiple, interleaved threads. Their GRE algorithm is motivated by the need to make a correct decision about whether an NP should be pronominal, reduced, or full, a decision that they claim should be based on three factors: (a) distance of the target referent from its nearest antecedent; (b) whether the last mention of the referent was in the same thread as the current utterances; and (c) the likelihood that the pronoun will be resolved correctly.

Perhaps the most detailed account of the dependency of GRE on discourse context is Krahmer and Theune's (2002) extension of the IA. The authors define a definite description as *suitable with respect to a referent r* if that referent is the most salient entity of its TYPE in the immediate context.[14] This is an identical strategy to that adopted by Passonneau (1995), whose algorithm attempts to distinguish a referent from other entities of the same type in the current utterance context.

Based on an empirical study, salience is defined by Krahmer and Theune as a combination of grammatical function, as per the Grosz et al. system, and Hajiĉova's hierarchy of salience. This is the basic mechanism which determines the distractors from which an intended referent is to be distinguished, and potentially reduces the processing required for the content determination process, which otherwise depends on a PO as per the original algorithm. Another innovation is the integration of syntactic formulation with content determination. The Krahmer-Theune algorithm maintains a description not as a set of properties, but as an expanding syntactic tree. Part of the reason for this is that, since the suitability of a definite is determined by its salience, failure to distinguish a referent from its most salient distractors will result in an indefinite NP. The algorithm itself is not explicit about the process of syntactic construction; however, it follows a general trend in this area of GRE, namely, to introduce aspects of realisation where these are motivated by discourse considerations.

As the preceding discussion makes clear, most algorithms in this area maintain the assumptions introduced in the earlier problem definition, but take into account whether a referent is in focus, in order to reduce the number of distractors that the referent is to be distinguished from. An exception to this general trend is the version of the Greedy algorithm proposed by Siddharthan and Copestake (2004). Their concern is with the use of GRE in applications, such as text simplification or summarisation, where the input is not a well-defined Knowledge Base, but a text. Given the assumption of textual input, any discourse entity is by definition one which has been introduced already. In producing a reference, Siddharthan and Copestake's algorithm seeks to maximise the distinctiveness of an entity relative to its distractors in context, by taking the semantic properties

[14] Note that the reference to TYPE assumes, in line with Dale and Reiter (1995) and previous psycholinguistic research, that the object class is the basic device by which an entity is recognised and processed.

predicated of the entity into account and seeking, among the lexical items known to be true of the entity, those which are conceptually furthest from the lexical items known to be true of its distractors. The notion of conceptual distance is operationalised in terms of lexical relations relations in WordNet. For example, if an entity has a property which is antonymous to that known to hold of a distractor, this increases the 'discriminatory value' of the property. This approach to greed in GRE is distinguished by its explicitly taking lexicalisation into account. In so doing, it combines the extensional success criterion of GRE (properties must be true of an intended referent, and only of it) with a consideration of lexical semantics and realisation. However, in using WordNet, it can in principle run into problems, since lexical relations in this database are defined between *senses*, not words as such. Thus, sense ambiguity can pose a very real problem.

This approach has some relationship to the lexically-driven content determination algorithms discussed later in this thesis (see especially Chapter 7). However, the focus in these algorithms is not on maximising semantic distance from distractors, but in maximising semantic closeness between elements of a set of intended referents.

To summarise, considerations of the discourse context have led to extensions of the GRE in two directions. In line with its Gricean motivations, distinguishing a referent from distractors in its immediate context potentially reduces overspecification. The algorithms proposed in this area have also blurred some of the boundaries between text planning, content determination and realisation in GRE, due to strong dependencies between the form of a referring expression (including its definiteness and whether or not it is pronominalised), and the accessibility or salience of the entity it identifies. These proposals represent ways of bridging potential gaps between text planning, content selection and realisation.

2.7.2 Expressiveness: Relations and gradable properties

As described so far, algorithms such as the IA and its predecessors are fairly limited in scope, since they limit content determination to literals – that is, one-place predicates. Moreover, in their original formulation, it was invariably assumed that the KB contains *crisp* properties, that is, properties which either hold true of an entity or not. Thus, none of the algorithms discussed so far would be capable of generating a reference to e_2 in Table 2.1 in terms of its POSITION, because this would require taking into account its spatial relationship to e_3, to yield, for example, *the woman to the left of the princess*. Position could actually be conceptualised differently. It so happens that all of the figures in the painting represented in Table 2.1 are standing in the foreground, forming a rough line from left to right. It is possible, then, to speak of the location of an entity in this domain in terms of how far left or right it is. To do this, however, the notion of *left* or *right* would have to be considered gradable, that is, a GRE algorithm would have to be able to handle the notion that something is *more or less* to the left or right. This section takes a closer look at some proposals to deal with these kinds of properties; it will also provide a further opportunity to consider the impact of the IA on GRE, since most of these proposals take it as their starting point. In so doing, the methods discussed highlight a lacuna in current research, namely, a general lack of solid empirical evaluation of the algorithms proposed, which are often based on previous psycholinguistic work, but stop short of testing its algorithmic interpretation.

2.7.3 GRE with relational properties

If n-ary predicates are brought into the picture, the GRE task becomes more complex. The property $\langle$POSITION : *left_of(e_3)*$\rangle$ introduces a relatum, e_3, which also needs to be described, and presumably also distinguished. Talking of e_2 in Table 2.1 as *the woman left of the woman* is potentially misleading (all the women in the table except for the rightmost one are standing to the left of a woman). Thus, a description could potentially become very complex, with relata being described in terms of other relata. A different sort of problem occurs when the only way to distinguish two entities is in relation to one another. In this case, the relatum would have to be distinguished in relation to the referent, which in turn has to be distinguished in relation to the relatum, and the descriptive process can become infinitely recursive (cf. Novak, 1988, for an early discussion of this problem).

These two issues were the main motivations for an early, constraint-based algorithm for the generation of relational descriptions by Dale and Haddock, proposed as an extension to Dale's (1989) GR algorithm. The algorithm maintains a constraint network $N = \langle D, C \rangle$, where D (the description) is a constraint set, initially empty, and C is the set of sets corresponding to the entities denoted by each property in D. The goal of the algorithm is to populate D with properties such that $\bigcap_{S \in C} S = \{r\}$. Relations are handled by maintaining a stack of descriptive goals, initially containing only the goal to refer to r, the intended referent. If a property is added to D which contains a constant (a KB identifier in the current terminology), then a further goal is pushed onto the stack to identify this relatum. Thus, if the algorithm were to select $\langle$POSITION : *left_of(e_3)*$\rangle$ in the process of describing e_2, e_3 would find itself on the stack, awaiting a description. Therefore, the generation of relational descriptions is conceived of in terms of a tree of sub-goals, each of which contributes to the final description of r. Infinite recursion arises when sub-goals are introduced to refer to entities already on the stack, the algorithmic correlate of the scenario in which the initial goal of describing e_2 results in a sub-goal to describe e_3, which in turn introduces the new goal to describe e_2. Dale and Haddock's solution was to impose a heuristic whereby no information in the KB can be presented more than once in a description. In this way, once the property $\langle$POSITION : *left_of(e_3)*$\rangle$ is considered, the property $\langle$POSITION : *right_of(e_2)*$\rangle$ cannot be re-used.[15]

Since the work of Dale and Haddock, approaches to relational descriptions have taken infinite recursion and descriptive complexity as a crucial aspect of the problem (e.g. Horacek, 1997; Varges, 2004). However, the impact of the IA has been to extend considerations of 'preference' to relational properties. In a recent model by Kelleher and Kruijff (2006), which uses the IA to generate relational locative expressions, it is assumed that relations are cognitively costly, and should be ordered towards the end of the PO. Furthermore, spatial relations are ordered with respect to each other, so that topological relations such as *near* are preferred over projective relations such as *above*, which describe regions in a particular direction from an object (Kelleher et al., 2006). Further, Kelleher and Kruijff (2006) argue that even within these two broad categories, further ordering can be imposed on the basis of psycholinguistic principles. For example, relations in the vertical dimension are preferred over horizontal ones (cf. Bryant et al., 1992; Gapp, 1995; Arts, 2004).

[15] Obviously, this is assuming that the algorithm 'knows' that one property is inferrable from the other.

In this approach, then, relations are not added if they are sufficiently discriminatory; rather, the view is that they should be avoided unless unless absolutely required. However, this algorithm still maintains an important aspect of the Dale and Haddock algorithm, namely that it includes sub-goals to describe relata, called 'landmarks' by Kelleher and Kruijff. These are handled by subsidiary calls to the IA. Infinite recursion is avoided because such a call is preceded by a tripartite division of the context set into (a) target referent, (b) distractors, and (c) candidate landmarks, described as those entities which are not identical to the target, and are not in the distractor set. The distinction between the distractors and candidate landmarks is possible because of a distinction between different kinds of spatial relations. Since the IA iterates through these relation types in a predetermined order, a subsidiary call to describe a landmark (relatum) of the target need only focus on those objects which stand in that relation to the target, so that everything else is a distractor.

This approach shows how a theoretically-motivated ordering among attributes can reduce not only algorithmic, but also descriptive complexity in GRE. However, the proposal shares with most other approaches a lack of empirical backing. To be sure, there is significant psycholinguistic evidence for the sorts of orderings proposed by the authors. Yet, the algorithm itself is frequently not evaluated, which leaves open the possibility that its procedure is not an accurate reflection of the relevant human tendencies. More than a criticism of this specific approach, this is an objection that can be levelled at most existing work in the field, where a common approach is to adopt psycholinguistic generalisations, but to stop short of evaluating their algorithmic interpretation. It is only recently that exceptions to this rule have begin to appear (e.g. Jordan and Walker, 2000, 2005; Gupta and Stent, 2005; Viethen and Dale, 2006). These recent empirical studies – which all focused on the IA to some extent – are reviewed in Chapter 4, which presents a new evaluation against a corpus which was purposely designed to address the particularities of the GRE content determination task. The corpus itself is described in the following chapter.

2.7.4 Gradable properties

Another recent development in the GRE literature, which also relinquishes some simplifying assumptions about Knowledge Representation, involves gradable properties (van Deemter, 2000, 2006). The properties of interest are *dimensional*, those that Bierwisch has characterised as corresponding to 'quantitative evaluations regarding dimensions or features' (Bierwisch, 1989, p.71), and which are often realised in natural language as adjectives that involve comparison between a contextually delimited set of entities.[16] Thus, values of an attribute such as SIZE place entities on a scale, and the assignment of a property such as $\langle$SIZE : *large*$\rangle$ to an entity in a particular context involves a comparison between that entity and others.

van Deemter's starting point is a semantics for definite descriptions containing gradable properties, whereby an expression such as *the large (n) T* is taken to mean *the largest (n) entities of type T*. Some evidence that this is indeed how people interpret definite descriptions was cited earlier in §2.6.2 (p. 42), where Sedivy et al. (1999) found that expressions such as *the tall cup* were interpreted based on a comparison of objects in the immediate visual context, and defaulted to the

[16] Bierwisch contrasts dimensional properties to *evaluative* properties, such as *clever* or *smart*, which, although vague or gradable like dimensional ones, often presuppose a global standard of comparison, rather than a purely contextual one. For example, the statement *x is clever*, on at least one of its interpretations, presupposes a scale of *cleverness* on which *x* scores high.

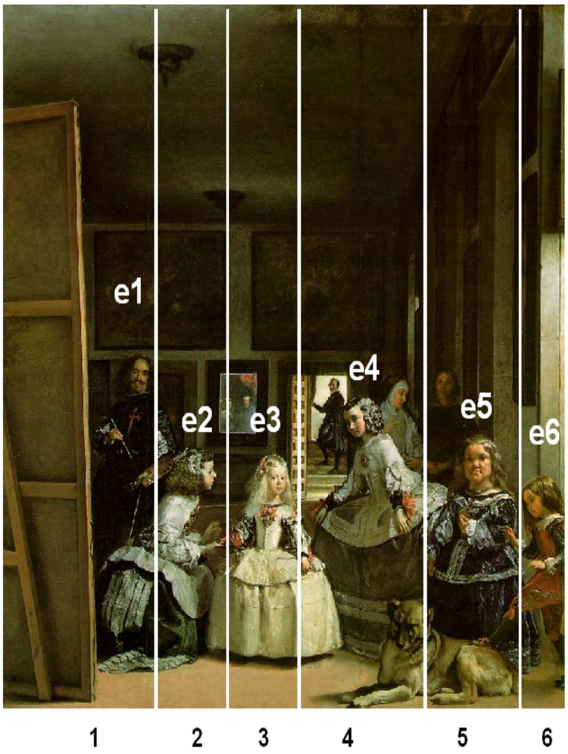


Figure 2.2: The painting, with overlaid grid

entity of the relevant type that had the largest relevant value. The proposal for the treatment of gradability in GRE is based on the assumption that the properties of interest are represented in the KB as numeric-valued attributes, of the form $\langle \text{A} = n \rangle$. To continue with the example introduced in the previous section, suppose that in addition to the relational attribute POSITION, the domain in Table 2.1 also had a numeric-valued attribute – call it X-DIMENSION – represented as a numerical coordinate expressing the position of an entity in the painting from its leftmost edge, as shown in Figure 2.2. e_2 can therefore be said to have the property $\langle \text{X-DIMENSION} = 2 \rangle$. She can therefore be described as *the leftmost woman*. Note that the expression *leftmost* presupposes a comparison between the relevant entities (in this case, the other women in the painting), on the basis of how far to the left they are. To achieve this, a GRE algorithm would have to deal with numeric-valued attributes in a way which inferred relationships of comparison between entities. In van Deemter's framework, this is achieved by an algorithm that is divided into three main parts:

1. A compilation step in which the KB is transformed to yield a new, explicitly comparative numeric representation of gradable attributes, using a simple form of inference;

2. The content-determination proper, based on an extension of the Incremental Algorithm;

3. A post-processing step in which further inference rules are applied to the output of the previous step.

The first, pre-processing step compiles every property of the form $\langle \text{A} = n \rangle$ into inequalities. For example $\langle \text{X-DIMENSION} = \mathit{2} \rangle$ is transformed into $\langle \text{X-DIMENSION} > \mathit{1} \rangle$, $\langle \text{X-DIMENSION} < \mathit{3} \rangle$, $\langle \text{X-DIMENSION} < \mathit{4} \rangle$, and so on. The outcome of this process is a new KB in which the original attribute has been 'compiled out' into two new attributes, corresponding to a 'greater-than' and a 'smaller-than' comparison. The second, content-determination step involves incorporating gradables into the IA. van Deemter's proposal is that such properties should be placed at the bottom of the preference order, a proposal which is compatible with the research cited earlier (e.g. Belke and Meyer, 2002) suggesting that relative or gradable attributes have lower codability (are less preferred and more likely to be filtered out unless absolutely required) than absolute ones, because of the cognitive cost involved in carrying out comparison between domain entities (but cf. Chapter 3). Placing gradable properties at the bottom of the preference order also has the consequence of increasing the chances that the domain of applicability of such properties is incrementally circumscribed. For example, assume that the algorithm has the preference order $\langle \text{TYPE}, \text{X-DIMENSION} \rangle$. If the IA selects TYPE and then selects the gradable attribute, the interpretation of the description, under the semantics proposed by van Deemter, will involve a comparison only between those entities which have the selected values of TYPE (thus, how far left e_2 is is judged in relation to how far left the other women in the painting are).

However, gradable properties are also ordered with respect to each other: the proposal is to always consider the property expressing the largest difference. Once again, this is in part based on psycholinguistic evidence by Hermann and Deutsch (1976, also cited in Levelt, 1989). In their experiment, they found that subjects who were asked to describe candles varying in height and width tended to select the attribute which expressed the largest difference. Thus, given a target referent which was significantly taller than other candles in the context, but only slightly wider, subjects were far more likely to include the HEIGHT rather than WIDTH in their descriptions. This is achieved by ordering inequalities on the basis of *logical strength*.[17] For instance, the logically strongest inequality of the form $\langle \text{X-DIMENSION} < n \rangle$ which is true of e_2 is the one with the smallest value of n, that is $\langle \text{X-DIMENSION} < \mathit{3} \rangle$.

The final, post-processing step involves carrying out some further inference on the description returned by the IA. If a description is returned which contains an inequality $\langle \text{A} > n' \rangle$, and n' is the largest of all the values of A in the KB which are true of the entities denoted by the description, the property is replaced with one expressing the maximality of the value directly. Thus, the description $\langle \text{TYPE} : \mathit{woman} \rangle \wedge \langle \text{X-DIMENSION} < \mathit{3} \rangle$ becomes *the leftmost woman*. This is compatible with the semantics of descriptions containing gradable properties, where the definite description is always interpreted as denoting the entities with the most extreme value of the relevant gradable attribute, unless it specifies otherwise.

A second kind of inference is carried out to correct a possible undesired outcome of the IA as described. Compiling numeric-valued attributes into two inequalities ($>$ and $<$) may result in a description which contains two properties $\langle \text{A} < n' \rangle$ and $\langle \text{A} > n'' \rangle$, where only one absolute value in the original KB falls in the interval (n', n''). In this case, the two inequalities are replaced by the original, absolute-valued property.

[17] A property p is logically stronger than q iff $p \models q$ but the reverse does not hold.

It could be argued that the inference carried out by this algorithm is motivated by a Gricean concern to 'say no more than is required'. This is already evident in the conversion of absolute values into relative inequalities, based on the assumption that precise numeric values should be avoided, at least in everyday communication, because they are too informative. It is also evident in the preference given to logically stronger properties, which have the potential consequence of minimising the size of a description, since a property of the form $\langle \text{A} > n \rangle$ will subsume all properties of the form $\langle \text{A} = m \rangle$ where $m < n$. Similarly, replacing inequalities with 'superlative' properties whenever possible is a way of reducing the information content of a description, by marking the extremity of a value without directly stating a number.

Like the proposals on anaphora discussed earlier in §2.7.1, van Deemter also extends his approach beyond 'pure' content determination, making some proposals for the realisation of descriptions containing gradables, focusing especially on when to use a base form for a dimensional adjective, and when to use superlative or comparative forms. Based on the semantics proposed for gradable properties in definite descriptions, the basic idea is that base-forms are interpreted as extreme values, just as superlatives are. This is expected to interact with the size of the 'gap' between the extreme value and the next in the KB (cf. van Deemter, 2004, for an empirical evaluation of these claims). The difference between this approach to realisation and those cited earlier, such as Krahmer and Theune (2002) (cf. §2.7.7 below), is that this proposal argues for *not* interleaving content determination and realisation. Given that a description can contain two inequalities corresponding to values of the same attribute (for example $\langle \text{SIZE} > n' \rangle$ and $\langle \text{SIZE} < n'' \rangle$), the realisation of such a description may depend on taking both inequalities into account *together*, yielding for example *the leftmost woman but one*. In an incremental content determination procedure like the IA, where these attributes are considered separately for inclusion (recall that the two inequalities are considered to be different attributes once the compilation step has been carried out), it is not possible to do this.

This proposal for gradability will resurface in Chapter 4, where it is used in an evaluation study of the algorithms discussed in this chapter, against a corpus of human-authored descriptions.

2.7.5 Logical completeness and plurality

In the problem definition of §2.4, the focus was on reference to individuals (i.e. singular reference). However, plural reference is commonplace in any NL discourse. By our original definitions, no GRE algorithm can generate a reference to a set. Indeed, it is only quite recently that authors have begun to focus on plurality and the novel problems, both logical and empirical, that it introduces. This has also been the case in the psycholinguistic literature. While a wealth of research exists on singular reference, plurals have rarely featured in the experimental paradigms discussed above. The psycholinguistic research on plurals (on which see Chapter 6) has focused mostly on plural anaphoric pronouns. The exception is a few early studies that focused on the relative cost of interpreting plurals compared to singulars.

At this point, it is worth introducing a new distinction, between what I will call **disjunctive** and **non-disjunctive** descriptions. In a non-disjunctive description, properties are logically conjoined, and the extension of a description is determined via set intersection. For instance, example (2.1) could be represented as in (2.17). In case there were two or more entities satisfying this description in the domain, this logical form would no longer be true of its intended referent (e_3)

but would be true of the set as a whole. Linguistically, the difference would be that the head noun would be morphologically plural. So far, the implicit assumption in this discussion has been that all descriptions returned by an algorithm are non-disjunctive; this is also the underlying assumption of Definition 3. On the other hand, suppose a reference to $\{e_6, e_7\}$ in Table 2.1 were required. One possibility is *the dog and the girl*. Here, the logical form would take the form of a disjunction (2.18), and its extension is determined via the union of $[\![\langle \textsc{type} : \textit{girl} \rangle]\!]$ and $[\![\langle \textsc{type} : \textit{dog} \rangle]\!]$.

(2.17) $\langle \textsc{type} : \textit{woman} \rangle \wedge \langle \textsc{clothing} : \textit{wears\_white} \rangle$ (=2.1)

(2.18) $\langle \textsc{type} : \textit{girl} \rangle \vee \langle \textsc{type} : \textit{dog} \rangle$

Note that what I have called a disjunctive description (2.18) corresponds to a linguistic **coordination**.[18] Extending the standard GRE algorithms to deal with conjunctive plurals is fairly straightforward; the only requirement is to make our intended referent a set (R instead of r). An algorithm would then check, for a given property, whether the *set* of intended referents is a *subset* of its extension. This is the essence of van Deemter's (2000) IA$_{plur}$, a version of the IA which handles plurals. Case (2.18) is a little more complicated, because no algorithm that produces exclusively conjunctive descriptions can handle it. In other words, IA and its predecessors are logically incomplete insofar as they cannot guarantee that a description will be found for a set of referents whenever one exists (van Deemter, 2002). Another aspect of this problem involves **negation**. In our domain, it is possible to refer to e_7 as *the dog*, but also (somewhat pedantically) as *the only non-human figure*. It is however possible to construe domains in which the only way to refer to an intended referent is by negating a property. In general, whenever there are two entities a and b such that $\mathcal{P}_b \subset \mathcal{P}_a$, then the only way to distinguish b from a is to negate a property in $\mathcal{P}_a - \mathcal{P}_b$.

In order to handle negation, an algorithm would need to calculate the complement of a property. This is the first innovation in IA$_{bool}$ (van Deemter, 2002), a version of the IA with full Boolean completeness. By the Closed World Assumption made in §2.4 (p.31), the negation of any property is assumed to be true of an entity if that entity is not explicitly listed as having that property in the KB. van Deemter proposed the addition of an initial stage to the IA in which negations are calculated by taking the complement of properties in the KB and adding them to the PO.

The other feature of IA$_{bool}$ is its handling of logical disjunction. Van Deemter's proposal is based on the observation that for any set of referents, if there exists a partition of that set such that every element of the partition can be conjunctively identified, then a disjunctive description exists for that set. In line with this observation, IA$_{bool}$ proceeds in a stepwise fashion. Step 1 is the original IA, which attempts to find a conjunctive description for R. If that fails, then the algorithm proceeds by considering disjunctions of properties of increasing length, conjoining them to the description as before. This means that IA$_{bool}$ searches through combinations of disjunctions of length k, for increasing values of k. The output of IA$_{bool}$ is a description in Conjunctive Normal Form (CNF). These innovations require only two changes to Definiton 2, whereas the success criterion defined in Definition 3 remains the same.

Definition 4. GRE Problem Instance (Revised)
A GRE problem instance is a 4-tuple $\langle K, R, P_R, D \rangle$ where:

[18] Here and in what follows, the term *coordination* is used for linguistic constructs such as *and* and *or*.

- $K = \langle U, \mathbb{P} \rangle$ is a KB;
- $R \subseteq U$ is a set of intended referents;
- $P_R \subseteq \mathcal{P}(\mathbb{P} - \emptyset)$ is a set of sets of properties ('disjunctions') such that:
    - $\forall P \in P_R : R \subseteq \left(\bigcup_{p \in P} [\![\, p \,]\!] \right)$
- $D \subseteq P_R$ is a *description*, the set of properties selected to describe R.

By this new definition, P_R, the set of relevant properties in $\mathbb{P}$, is a set of sets or disjunctions. This takes into account van Deemter's extension, since the search space of a GRE algorithm is no longer populated exclusively by literals (which are now the singleton sets in P_R), but also by combinations of these. Although this proposal was made in relation to the IA, it is easy to see how the other algorithms in §2.5 could be extended in the same way. Indeed, the only changes that need to be made to IA, FB and GR as formalised in previous sections is an additional step which, having tested a property for inclusion in a description, enqueues disjunctions involving that property. In the case of FB, something similar was already happening with conjunctions, since this algorithm had to search exhaustively. However, introducing this extension implies that the ordering relation $>>_{p_x}$, which determines the next property to be visited by the algorithm, now has to be generalised to deal with comparisons between disjunctive and non-disjunctive formulae. As discussed below, most authors in the literature on plurals have assumed that communicative economy has a role to play in this case too: excessive length and logical complexity are to be avoided, and therefore disjunctions (and negation) should be dispreferred relative to non-disjunctive combinations.

An alternative to IA$_{bool}$ (van Deemter and Halldórsson, 2001) uses **satellite sets** to describe a set of referents. The satellite set $sat(e)$ of a domain entity e is defined as follows:

$$\forall e \in U : sat(e) = \bigcap_{p \in P_e} [\![\, p \,]\!] \tag{2.19}$$

The authors propose an algorithm that calculates the satellite set for each entity in the domain as a first step. The process of finding a description for R boils down to checking whether the union of satellite sets of each $r \in R$ is equal to R. Once again, adding negation is a relatively easy step.

Both of these algorithms provably achieve Boolean completeness. More precisely, this can be proven if the properties in the KB are **distributive**. Simplifying somewhat, distributive predicates are true of each individual element of a plurality. Thus, in our running example, the property $\langle$POSTURE : *standing*$\rangle$ is distributively true of $\{e_3, e_4, e_5, e_6\}$. Collective predicates, such as *meet*, are true of a plurality as a whole. In order to handle these under the standard framework, algorithms would require a richer representational formalism, in which it would be possible to specify that the extension of such properties is a collectivity. An alternative approach is that proposed by Stone (2000), whose constraint-based algorithm searches for a salient cover of a set of referents, and does not require the assumption that all properties in the KB are distributive.

The logical completeness of IA$_{bool}$ comes at a cost. First, disjunctions cause a combinatorial explosion in the search space. In particular, letting $n_p = |\mathbb{P}|$, and assuming that the resulting description is of length n_d, there are $\binom{n_p}{n_d}$ ways of choosing n_d properties from the available set (van Deemter, 2002). This increases by a factor of 2 if negations are taken into account. The

Satellite Sets Algorithm resolves this problem somewhat (especially if the construction of satellite sets is assumed to be an offline step), but trades off on incrementality. The same argument applies, mutatis mutandis, to GR and FB, if extended as indicated above.

2.7.6 Plurals and naturalness: The return of Brevity

The concern about computational cost in the GRE literature probably stems from Dale and Reiter's original motivations for the design of IA. Arguably, however, theoretical complexity is secondary to the ultimate desideratum of GRE (and NLG in general), namely, naturalness of output. Yet it is far from clear that either of the two algorithms described here achieve this. Originally proposed as a better reflection of the psycholinguistic data, IA formalises the notion of preference and accounts for redundancy by using an ordered list of attributes. In IA$_{bool}$, the preference order becomes a much vaguer construct, since after the initial phase (the original IA, perhaps with negation), it becomes quite difficult to determine in what sense a disjunction should be ordered before another of the same length. Another problem is naturalness. It is relatively easy to construct domains in which IA$_{bool}$ yields descriptions of significant logical complexity, when simpler ones exist (cf. Gardent, 2002; Horacek, 2004). Part of the problem here is what van Deemter calls the *double incrementality* of the algorithm, 'double' because it performs gradient descent on literals, and then again on Boolean combinations. For example, suppose $R = \{e_4, e_5\}$ in our domain. During its first pass, the algorithm selects *woman*, which excludes e_1 and e_7 and is true of R. Next, it selects *maid*, which excludes e_3. During its next phase, the disjunction *wears_brown* $\vee$ *wears_black* is selected, and the process terminates with the description in (2.20).

(2.20) $\langle$TYPE : *woman*$\rangle \wedge \langle$ROLE : *maid*$\rangle \wedge (\langle$CLOTHING : *wears_brown*$\rangle \vee \langle$CLOTHING : *wears_black*$\rangle)$

In a sense, this procedure takes the Incremental Algorithm to its limit. In so doing, I believe it highlights a mismatch between this notion of incrementality, and the notion that psycholinguists have appealed to in explaining overspecification. Wundt's Principle implies that the processes of perception, conceptual formulation and realisation are closely coupled in time, something that also emerges from online studies of production and comprehension. In case a description such as (2.20) were required, a speaker following IA$_{bool}$ as a strategy would have to consider successive combinations of properties of both referents in tandem, hardly plausible given the preceding overview of the evidence. This is not meant as a critique of IA$_{bool}$ itself (which was not claimed to reflect psycholinguistic tendencies). It does however highlight one possible reason for the observed lack of naturalness of its output in certain cases. Another issue that is worth raising is the lack of linguistic transparency in the logical form. How is a description such as (2.20) to be realised? A direct rendition of the logical form into an NP is of course possible, but in the case of complex logical forms, the outcome could potentially be unnatural. It could be argued that realisation is in fact a problem to be dealt with elsewhere. However, as we saw in §2.7.1, some GRE tasks do require a consideration of more linguistic problems. In the case of plurals, some authors have proposed similar extensions.

Two further developments were proposed in the literature on generating plurals. Gardent (Gardent, 2002; Amoia et al., 2002) returned to the view espoused in the Full Brevity algorithm, with a constraint-based approach that uses set-cardinality constraints to find the briefest possible description of a set (cf. Gardent et al., 2004). By contrast, Horacek (2003, 2004) proposed a best-first search procedure, which addresses three limitations of IA applied to Booleans. The first two of these – excessive redundancy and logical complexity – are highly related. The third is based on a criticism of the IA which assumes that descriptions are constructed by conjunctions of properties, hence are represented in Conjunctive Normal Form (CNF). Like IA$_{bool}$, Horacek's proposal is a generalisation of the IA. However, it includes the following three types of constraints, aspects of which had been introduced in an earlier algorithm focusing on the generation of descriptions containing n-ary relations (Horacek, 1997):

1. A preference for descriptions involving explicit exclusion of distractors (*the cars which are not red*), rather than exhaustive description of referents (*the blue car and the green car and* . . .), if the former results in a briefer description;

2. Exclusion of logical forms that are difficult to realise in NL, without significant ambiguity. The algorithm attempts to identify a set by partitioning the set of intended referents into subsets, and describing these subsets;

3. An abandonment of the idea that a referring act necessarily consists of a single referring expression. If a single expression becomes too lengthy, the algorithm generates a sequence of such expressions. This requires an *a priori* threshold for the complexity of a description.

Perhaps the most interesting of these innovations from the algorithmic point of view is the idea of partitioning, also adopted by van Deemter and Krahmer (2006) in the graph-based framework. Their partitioning algorithm begins by attempting to run IA on the input set R, failing which, R is repeatedly $k-$partitioned, for values of k up to $|R|$, until a division is found such that every element of the partition can be distinguished.

Partitioning results in a division of labour in the process of referring to a set, and also changes the 'naively incremental' behaviour of IA$_{bool}$. While van Deemter and Krahmer's algorithm focuses on the semantics of descriptions, Horacek's linguistically-oriented mechanisms, blurring the distinction between content determination and realisation, offer a promising way to ensure greater transparency in the mapping from logical forms to NL representations. This approach raises several interesting questions about the relationship between different microplanning and realisation tasks. For instance, the decision of whether to generate a single referring expression or several is essentially an aggregation decision. Expressibility constraints, such as the avoidance of overlengthy descriptions and the exclusion of logical forms that cannot be straightforwardly mapped to NL representations, is a way of dealing with a potential generation gap (cf. Meteer, 1991), by making content determination more linguistically driven. This is part of a broader trend in the NLG literature, discussed in §2.7.7.

What is lacking in these constraints is an empirical foundation. Many of the issues raised by both Gardent and Horacek are well-taken. It is undesirable for an algorithm to produce a logically very complex expression. Murphy (1984), in one of the rare psycholinguistic studies on full NP plural reference, reported significantly longer reading times for disjunctive descriptions (i.e. those involving linguistic coordination), compared to conjunctives, suggesting that highly complex NPs would be even more difficult to process. Similarly, there is some research suggesting that logical operators incur significant interpretive expense on the part of listeners. Johnson-Laird (1983) proposed that this was because disjunction requires listeners to construct and juxtapose multiple mental models. Suppes (1971) reported a developmental study, broadly falling within the Reference Task paradigm, in which children had to interpret references containing conjunctions, disjunctions, and/or negations. Performance on instructions containing the latter two types was consistently worse.

Despite the prima facie validity of the hypotheses discussed here, it is worth reiterating that very little in the way of systematic empirical investigation and/or evaluation has been done. Part of the problem is that, with the exceptions noted in this section, psycholinguistic studies have focused exclusively on the singleton case, perhaps assuming that the results carry over quite naturally to the plural case. Be that as it may, disjunction (and negation) introduce novel problems that warrant empirical investigation. As in the case of context-sensitivity, authors in this field have also begun to raise questions about the coupling of content determination and other aspects of NLG.

2.7.7 GRE and 'global' approaches to generation

The proposals for tighter integration of GRE with realisation and discourse or text planning emerged from extensions of the traditional domain of referring expressions generation. The new problems that arise with context-sensitivity and logical completeness echo some of the broader concerns in the generation literature to do with Logical Form Equivalence (LFE; Appelt, 1987a) and the Generation Gap (Meteer, 1991) – problems which arise due to a mismatch between what is planned at a strategic level by a computationally autonomous, linguistically independent planner, and what the tactical component can handle further downstream.

As first formulated by Appelt, LFE deals with the desirability of having a symmetry between a component which deals with semantic forms, and one which deals with their NL realisation. As Shieber (1993) noted, this problem only exists to the extent that the logico-semantic ('strategic') and the linguistic ('tactical') components are taken to be separate, that is, the representation of meaning takes place without any input from the linguistic module. The Generation Gap of Meteer (1991) could be viewed as a consequence of this separation, whereby a strategic component plans an utterance to satisfy a given goal, while a tactical component lacks the resources to realise this goal linguistically. These issues are relevant to NLG, not only because they echo some of the concerns outlined earlier in relation to extensional equivalence[19] (cf. §2.4), but because of the way the content determination problem has traditionally been defined in this area. As the preceding discussion sought to make clear, GRE is a semantically intensive task, and most work in the area has formulated solutions against the background assumption that 'the language is taken care of somewhere else'. It is only when expressiveness is extended, or context and Boolean completeness taken seriously into account, that possible mismatches appear between notions of adequacy defined at the semantic level, and their real linguistic outcome. Thus Horacek and Gardent's concerns with avoiding logical complexity, and making search more linguistically-informed, echo the argument made at the beginning of this chapter, to the effect that microplanning is something of a mixed bag, with content-determination problems existing side-by-side with, but informationally encapsulated from, lexicalisation and aggregation.

Some recent work in NLG has evinced a move towards an even more 'global' approach to

[19] This has indeed been characterised as a case of LFE; see (van Deemter and Halldórsson, 2001).

generation, where content determination is carried out opportunistically, depending on the available linguistic resources. Examples of this work include that by Barzilay and Lapata (2005, 2006) on the generation of reports of American football games from large databases. The central innovation here is the use of *collective* content selection, whereby several candidate elements for inclusion in a text are considered simultaneously, so that the coherence of the resulting text can be evaluated. This process combines content determination and aggregation. The collective content determination process ensures that the output text respects corpus-derived, domain-specific constraints about what content 'hangs coherently together'. In this way, the system avoids generating fragments that report on events that are unrelated. A somewhat different approach is the instance-based, overgeneration-and-ranking architecture of Varges and Mellish (2001). This system has two principal components: a grammar that (over-) generates sentences to express particular elements of content, and a comparison method that weighs the generated sentences against semantically annotated instances in a corpus. The crucial idea is to evaluate a sentence in terms of how well it expresses the given content, given similar instances in the corpus.

These trends in the literature arose as responses to the generation gap. They have been taken up to some degree in the GRE literature as well. Varges (2004, 2005b) proposed an extension of his instance-based generation architecture to referring expressions containing boolean operators and n-ary relations in a chart generation system. Several logical form fragments for a referring expression are produced by the rule-based system. The job of the chart generator is to realise these forms and combine them, with the search space being pruned of forms that cannot be adequately realised or combined with existing fragments, based on a corpus-derived evaluation metric. This approach is related to that of Horacek, since the content determination process is constrained by realisation and aggregation possibilities. However, it is more directly realisation-oriented.

A somewhat different approach is adopted in the description-building component of the SPUD system (Stone and Webber, 1998). At the basis of the system is a rich semantic representation, coupled to syntactic fragments (which are simple trees in a lexicalised Tree Adjoining Grammar). The semantics includes details of the contribution to the overall message that a fragment makes, and the requirements for the use of that fragment (e.g. what information it presupposes). Because of the close coupling of representations, the system plans syntax and semantics in tandem, keeping track of what parts of the overall communicative goal have been satisfied by the message constructed so far, and what new requirements are introduced by new parts of a description. The way a description is constructed is incremental, in the sense that new information is linguistically realised as soon as it is added, and the system's next actions are in part conditioned by the resulting state.

So far, research on global approaches to GRE has been somewhat disparate. Together with work on disjunctive descriptions and context, it shows a trend towards treating communicative intentions and surface-oriented constraints in tandem, constraining the semantics via the system's linguistic capabilities and its discourse plans.

2.8 Summary and outlook

This chapter began with an overview of GRE and its place in the NLG architecture. Some early approaches to the problem, framed within a theory of communicative action that takes into account intentions and goals, were shown to have made the wrong predictions. Subsequent developments paid more attention to the psycholinguistic literature, but maintained a largely semantic view of the GRE problem. Essentially, the view was that GRE was about selecting content for inclusion in a logical form. The problems of Logical Form Equivalence and the Generation Gap again reared their heads once these methods were extended to take into account contextual factors and logical completeness. As a result, recent developments have advocated a tighter coupling between semantic formulation and linguistic realisation.

To conclude this chapter, I will compare some of the work in psycholinguistics to these more recent models. The conclusion reached in §2.6 was that the way speakers achieve the prototypical communicative goal of a referring expression is not straightforwardly predictable from a model of communication that is framed at a purely intentional level. Rather, automatic processes related to the incremental selection of content and its realisation play a significant role in the final output. The centrality of automatic incremental processing is also evident in speaker-listener asymmetries. In GRE, Dale and Reiter's Incremental Algorithm gave new impetus to research on extending the capabilities of this component of NLG systems, advocating a stronger relationship between the achievement of communicative intentions, semantics, aggregation and syntax. This work has not tried to address psycholinguistic research in any degree of detail. I have argued above that empirical work has been lacking in this area in general, and that issues related to disjunctives and context would benefit from more empirical research. There are also some points of convergence between the psycholinguistic work reviewed and the recent computational literature on reference. One of these is the relativisation of communicative principles stemming from a purely intentional account, in order to explain or approximate speakers' tendencies. Another is the increased focus on how the achievement of intentions is modulated by linguistic and processing constraints. In the context of NLG, this is a way of addressing aspects of the generation gap.

This review has thrown up a number of gaps in the research on GRE. The first of these is a lack of empirical validation of hypotheses incorporated in algorithms, and subsequent evaluation. In the next chapter, I address this via an evaluation study that compares the output of the four main algorithms discussed here to human data. The study will also serve to draw attention to some linguistic features of reference, and will allow a more precise formulation of the three hypotheses outlined in Chapter 1. The subsequent chapters will take up a number of themes raised in this chapter, especially on the relationship between semantic and linguistic forces in NP generation. The focus will be primarily on lexicalisation and aggregation, and the algorithms proposed will be motivated by psycholinguistic and corpus-based work.

Chapter 3

Producing referring expressions: A corpus study

3.1 Introduction

It is clear from the overview in the preceding chapter that the current state of the art in GRE is dominated by models that build directly on the groundwork established by Dale and Reiter (1995). The Incremental Algorithm has served as a starting point for a number of later models, which have sought to extend the expressiveness and coverage of GRE (see, among many others Horacek, 1997; van Deemter, 2000, 2002, 2006; Kelleher and Kruijff, 2006, and the discussion in §2.7). In addition, the concerns that motivated the IA, especially computational efficiency, psycholinguistic plausibility, and success in achieving Gricean communicative effectiveness, have become central to developments in the field, making the IA a yardstick against which to compare other approaches (e.g. Gardent, 2002; Jordan and Walker, 2000, 2005).

I argued in Chapter 2 that this body of work lacks a sound empirical basis. No attempt has as yet been made to test Dale and Reiter's claim that the IA is superior to its predecessors on psycholinguistic grounds (apart from computational efficiency). Moreover, several later models which built directly on the IA – including those dealing with Booleans and gradable properties – have only been evaluated to a limited extent. This chapter and the next aim to contribute to filling this gap.

The rest of this chapter is structured as follows. I begin by reporting on a controlled experiment that resulted in the construction of the TUNA Corpus, a **semantically transparent corpus** of human-produced descriptions. §3.2 describes the rationale behind the corpus, its design, and the experimental method used for its construction. Following a description of the procedure (§3.4, p. 75), I go on to discuss the annotation scheme used (§3.5, p. 78) as well as an evaluation of the reliability of the scheme. The remainder of the chapter reports on an empirical investigation of the descriptions in the corpus, testing a number of hypotheses explicitly laid out in §3.3 (p. 73). The aims of this study are twofold. First, it is intended to serve as groundwork for an exhaustive evaluation of classic GRE algorithms, a topic which I turn to in the following chapter. Second, the study extends previous work in the psycholinguistics of reference, seeking evidence in the corpus of the following:

1. **The effect of communicative setting** (§3.7, p. 86): In the TUNA Corpus experiment, an attempt was made to manipulate the extent to which authors would perceive the setting

in which they were identifying objects as *fault-critical*, under the assumption that a fault-critical situation would increase the likelihood of overspecification. The results concerning hypothesis are not clear-cut, however. One reason may be that the experimental manipulation of this variable was too contrived. However, some evidence that the communicative task had some validity for participants is offered in §3.4.1 (p. 78). A more likely reason for the trends in the data is that the relevant experimental variable turned out to be somewhat confounded with another, namely, the possibility of referring to objects using their location.

2. **Attribute preferences** (§3.8, p. 89): The Gestalts Hypothesis explains people's tendency to overspecify with reference to mechanisms involved in the mental representation of objects, where certain attributes are more central than others. Therefore, descriptions should evince a higher likelihood of usage of these attributes *when they are not required for identification* compared to other attributes. Because of the way trials in the experiment were designed, it is possible to find such instances, as well as cases of underspecification, in which an attribute which is required to successfully identify an object is not included in a description. The results of this part of the study display a striking parallelism with results from previous studies, but also extend them by looking at a greater variety of attributes and also at underspecification.

3. **Extension of these findings to the plural case** (§3.9, p. 94): The TUNA Corpus contains plural descriptions. Since the plural trials in the experiment were in part constructed to address the Similarity Hypothesis laid out in Chapter 1 (which extends the Gestalts Hypothesis), this part of the study seeks empirical backing for this hypothesis. The main result is that the observations on over/underspecification and attribute preferences in reference carry over to the plural case, that is, people are very likely to overspecify with plurals, depending on the nature of the attributes involved. This is what an application of Occam's Razor would lead us to expect: the tendencies observed with singulars are also evident when a reference is made to a set rather than an individual. However, the data also shows that these tendencies are mediated by the similarity of the objects to which people refer. In particular, the tendencies observed in descriptions of two objects which share many of their perceptual attributes (e.g. are of the same colour, have the same size and face in the same direction) are roughly identical to those observed in the singular case, whereas they are less in evidence when the objects are perceptually dissimilar. Overall, this part of the study suggests that work on plural GRE which interpreted adequacy in terms of Gricean brevity and/or logical simplicity (§2.7.5, p. 58) was probably not on the right track. However, similarity is also playing a crucial role.

This analysis informs many of the decisions made in Chapter 4 for the evaluation of algorithms, especially in relation to the Incremental Algorithm. The second part of the thesis takes the investigation of the plural data even further.

3.2 Constructing a semantically transparent corpus of references

Microplanning tasks such as GRE, are **semantically intensive** because they mainly involve content determination. This makes existing corpora difficult to use for evaluation because such resources tend to consist of surface forms without semantic annotation. Thus, it is not always possible to tell whether a description in a text-only corpus is referential and what the communicative intentions underlying that reference were. Even if it were possible to find several descriptions of the same referential targets in such a corpus, individual differences in lexical choice and syntactic realisation may obscure similarities between them at the level of content. Thus, corpora texts are not ideal resources for evaluation; rather, it is necessary to have a transparent mapping from Natural Language to semantics. This is what the term **semantic transparency**, introduced by van Deemter et al. (2006), conveys. A semantically transparent corpus that is adequate for GRE as a content determination problem needs to satisfy the following requirements:

1. Provide (through annotation) a clear, normalised semantic representation associated with the linguistic content, which abstracts away from variations in lexicalisation where such variations do not directly impact the semantics.

2. Make the expressions in the corpus comparable.

3. Make explicit the domain representation against which such expressions were elicited.

Another desirable feature of a corpus for GRE evaluation is that it be **pragmatically transparent**, that is:

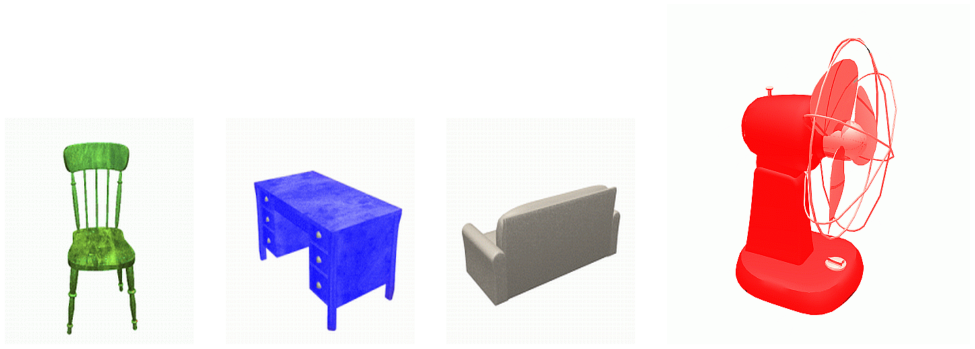
4. Control, as far as possible, the communicative intention as a result of which corpus expressions were produced, thereby minimising the risk of confounds due to intentions over and above those for which an algorithm was designed. In this way, human and algorithmic output can be compared on a more level playing field.

Though meeting this desideratum is perhaps less straightforward than meeting the other three, the present study also tried to restrict authors' communicative intentions to the goal of identifying objects, in line with the basic assumptions underlying many GRE algorithms.

The experiment reported here formed part of a larger collaborative study conducted as part of the TUNA Project[1]. The principal aim of the experiment was to collect a large sample of referential descriptions, against well-defined domains, which could serve both as a testbed in which to compare GRE algorithms in exactly the same domains – the topic of Chapter 4 – and as a set of experimental observations on which to test hypotheses about human referential communicative behaviour.

The method used was a variation on the classic reference task, in which participants were presented with successive trials, each containing a visual domain of objects, and were requested to identify a subset by typing a description. Participants were interacting remotely with a computer system, which gave limited feedback by removing some objects when a participant submitted a description. The choice of target referents was predefined in order to enable full control of what properties were minimally required in order to identify the set of referents. Throughout, it was made clear to participants that their principal aim was to *identify* the target unambiguously for the system they were interacting with. The corpus of data collected was then annotated at the semantic

[1] See `http://www.csd.abdn.ac.uk/research/tuna/`. The part of the study reported here, including design, implementation and data analysis, were the work of the present author. The other half of the study had a parallel design and an identical methodology, but used objects from a different domain of discourse.



TYPE	chair	desk	sofa	fan
COLOUR	green	blue	grey	red
ORIENTATION	forward	leftward	backward	rightward
SIZE	small	small	small	large

Figure 3.1: Example objects from the furniture domain

level, so that every description was paired to an explicit domain representation, and segments of every description were marked up to indicate which domain attributes had been used.

3.2.1 Materials

Referential domains, each corresponding to an experimental trial, were constructed using pictures of furniture and household items obtained from the Object Databank[2], a set of digitally created images developed by Michael Tarr at Brown University. The Databank consists of several pictures of everyday objects. For each object, six pictures are provided, representing the same object at six different orientation angles.

Four objects were selected from the Databank, corresponding to four values of the TYPE attribute. One of the main selection criteria for the objects was that each be easily recognisable, and that each have a clearly discernible front to facilitate the determination of its orientation. For each object, there were four versions corresponding to four different values of ORIENTATION. Pictures were manipulated using image processing software to create a version of each TYPE × ORIENTATION combination in four different values of COLOUR. Each resulting picture was also processed to yield two versions corresponding to SIZE: *large* pictures of 450 × 450 pixels, and *small* pictures of 250 × 250 pixels.

The ease with which participants recognised the type of the objects in their different orientations was determined on the basis of a pilot study involving 19 participants. Following this pilot, one value of TYPE (*television*) was removed, and exchanged for a different value (*fan*), because some participants had difficulty in discerning the type of object in some of its orientations.

Table 3.1 summarises the full set of attributes and values available in the final version of the domain. Examples of pictures representing objects with different attribute combinations are shown in Figure 3.1. The choice of attributes was motivated by a need to represent objects on

[2] `http://alpha.cog.brown.edu:8200/stimuli/objects/objectdatabank.zip/view`

COLOUR	TYPE	ORIENTATION	SIZE
blue	chair	forward	large
red	sofa	backward	small
green	desk	leftward	
grey	fan	rightward	

Table 3.1: Attributes and values in the*furniture* domain

qualitatively different dimensions. Psycholinguistic research has indicated that some attributes are 'preferred', in the sense that they get included in references irrespective of their contrastive value (cf. §2.6, p. 38). The set of attributes in this experiment provided a good variety. COLOUR, which has been found to be strongly preferred in experiments using the reference task, contrasts with SIZE, an attribute which is gradable, and hence has low codability, that is, its inclusion in the mental representation of an object incurs higher cognitive load because to determine the value of this attribute requires comparison to other objects (cf. §2.6.1, and Belke and Meyer, 2002). Although in the present case, the ratio between the two values of SIZE is reasonably large, making the size difference quite salient, Belke and Meyer showed that such properties tend to be filtered out unless they are contrastive.

As can be seen in Figure 3.1, values of ORIENTATION can be discriminated fairly easily. In the pilot study, none of the participants seemed to have problems with this attribute in descriptions where it was required, except for the one value of TYPE cited earlier, which was changed. However, there is the possibility that objects facing *left* or *right* incur more effort to be recognised and described. These orientations are not at a 90° degree angle, and object recognition in these cases requires mental rotation. In short, although the value of a target's ORIENTATION does not require comparison to other objects to be determined (unlike SIZE), it was included as an attribute which was expected to fall somewhere between COLOUR and SIZE in terms of preference.

3.2.2 Design

Twenty experimental trials were constructed, each consisting of one or two target referents, and six distractor objects. Trials were constructed such that the available attribute-value pairs in the domain were represented an approximately equal number of times, in order to avoid bias due to excessive familiarisation with certain properties. For example, of the 12 domains requiring ORIENTATION, a target faced*front* or *back* exactly half the time, and *left* or *right* in the rest.

Trials were counterbalanced, by taking into account the Minimal Description (MD) required to distinguish the targets. TYPE was never distinguishing in any trial, as it was assumed, based on robust psycholinguistic findings, that it would be included anyway. This is because, as explained in §2.6.1 (p. 38), the conceptual category to which an object belongs (especially the basic level value; cf. Rosch et al., 1976) is the primary dimension in the mental representation of the object. Thus, trials were constructed in such a way that TYPE wasn't required, either individually or in combination with other attributes, to identify the object(s).

For the remaining three attributes, there are seven possible unique combinations[3]. For each

[3] $\sum_{k=1}^{3}\binom{3}{k} = 7$

	Cardinality	Similarity	Minimal description	Singular example	Plural example
1	1 or 2	similar	COLOUR	the blue chair	the blue chair and desk
2	1 or 2	similar	ORIENTATION	the chair facing right	the chair facing right and the desk facing right
3	1 or 2	similar	SIZE	the large chair	the large objects
4	1 or 2	similar	COLOUR, SIZE	the large blue chair	the large blue furniture items
5	1 or 2	similar	ORIENTATION, SIZE	the large chair facing left	the large, left-facing chair and desk
6	1 or 2	similar	COLOUR, ORIENTATION	the blue chair facing left	the blue chair and fan facing left
7	1 or 2	similar	COLOUR, ORIENTATION, SIZE	the large blue chair facing left	the large blue desk and sofa facing left
8	2	dissimilar	COLOUR	N/A	the blue and red chairs
9	2	dissimilar	ORIENTATION	N/A	the chair facing left and the chair facing right
10	2	dissimilar	COLOUR, SIZE	N/A	the large blue chair and the small green chair
11	2	dissimilar	ORIENTATION, SIZE	N/A	the large chair facing forward and the small chair facing the back
12	2	dissimilar	COLOUR, ORIENTATION	N/A	the grey chair facing right and the green chair facing left
13	2	dissimilar	COLOUR, ORIENTATION, SIZE	N/A	the large grey chair facing right and the small green chair facing the back

Table 3.2: Within-subjects experimental design

one, a domain was constructed in which distractor objects were such that the MD combination was the minimal description for the target referents, and any successful reference would have to include at least the attributes in MD (unless LOCATION could be used; see below). Domain objects were placed within a 3 (row) ×5 (column) grid whose grid-lines were not visible. This grid structure was a sparse matrix, since out of the 15 cells, at most 8 (2 targets + 6 distractors) could be filled. The design consisted of one within-subjects factor, and one between-groups factor, as follows.

1. **Cardinality/Similarity (within; 3 levels)**: Trials were either **Singular**, containing exactly one referent, or **Plural**, containing two. Half the plural trials were **Similar**, that is, they had the same values of the attributes in MD. The other half were **Dissimilar**, with different values of the attributes in MD. Thus, if COLOUR was part of MD, one referent might be *blue* and the other *green*. In other words, MD in Dissimilar trials was a disjunction, whereas it was non-disjunctive in the Similar case. However, the two referents in a plural trial always had different basic-level values of TYPE, and always had identical values on the non-contrastive attributes (those not in MD). As an example, (3.1) below shows an example of the attribute-value pairs of the two target referents in these two conditions. Note that Similar referents were essentially identical, except that they belonged to different object classes. In both examples, MD consisted of COLOUR only.

(3.1) (a) (Similar)

$$\text{(a)}\ \Big\{\langle\textsc{type} : \textit{fan}\rangle, \langle\textsc{orientation} : \textit{right}\rangle, \langle\textsc{colour} : \textit{blue}\rangle, \langle\textsc{size} : \textit{small}\rangle\Big\}$$

$$\text{(b)}\ \Big\{\langle\textsc{type} : \textit{sofa}\rangle, \langle\textsc{orientation} : \textit{right}\rangle, \langle\textsc{colour} : \textit{blue}\rangle, \langle\textsc{size} : \textit{small}\rangle\Big\}$$

(b) (Dissimilar)

$$\text{(a)}\ \Big\{\langle\textsc{type} : \textit{fan}\rangle, \langle\textsc{orientation} : \textit{front}\rangle, \langle\textsc{colour} : \textit{blue}\rangle, \langle\textsc{size} : \textit{small}\rangle\Big\}$$

$$\text{(b)}\ \Big\{\langle\textsc{type} : \textit{sofa}\rangle, \langle\textsc{orientation} : \textit{front}\rangle, \langle\textsc{colour} : \textit{green}\rangle, \langle\textsc{size} : \textit{small}\rangle\Big\}$$

A possible shortcoming of this design is that having different-TYPE objects in the Similar condition may have biased authors towards producing disjunctive descriptions (consisting of at least two disjoint basic-level values of TYPE as in *the fan and the sofa*). However, as shown below, this did not deter Similarity from exerting an influence on people's content determination decisions. The issue of basic-level TYPE values in disjunctive descriptions will be treated in much greater detail in Chapter 5.

This within-subjects design yielded 20 trials. There is a gap in the experimental design, since there is no trial with MD corresponding to SIZE alone in the Dissimilar condition. This is because, since SIZE was a binary-valued attribute, it was impossible to construct a domain in which it would be uniquely distinguishing for two objects, with two different values. Examples of the different factor combinations are shown in Table 3.2 with some hypothetical target descriptions. Note that plural descriptions in the Similar condition can be non-disjunctive if either TYPE is omitted, or a superordinate (non-basic-level) TYPE like *furniture* is used.

2. **Communicative task (between; 3 levels)**: This factor manipulated whether the referential task was perceived as **Fault-Critical**. Half the participants were assigned to a version of the experiment defined as *fault-critical* (+FC), while the others were not (−FC). The main difference was in the instructions given to participants. In the +FC version, participants were told that the system they would be interacting with was being tested for use in situations were it was crucial that it understand linguistic messages correctly, and where no option to rectify errors would be available. The participants in the −FC group were not given this information, but were informed that they would have the opportunity to correct the system's mistakes if and when they arose. Thus, the latter condition allowed for repairs of referent identification failures. In addition, half the participants in the +FC condition were told that the system's domain representation was identical to theirs, in that it "knew" about the location of objects as perceived by the participants. The other half, as well as all participants in the −FC condition, were informed that the system had no way of telling how the domain objects were laid out on a participant's screen. Thus, one third of participants overall had LOCATION as a possible attribute (+LOC condition), in addition to those shown in Table 3.1. Because objects were placed in a 3 (row) × 5 (column) grid, this property was actually split into two attributes: X-DIMENSION (horizontal) and Y-DIMENSION (vertical), each of which took a numeric value. The three levels of Condition are summarised in Table 3.3. The full instructions corresponding for each condition are reproduced in Appendix A.

LOCATION was introduced in the +FC condition as a result of two observations made in the psycholinguistic literature. von Stutterheim et al. (1993) and Arts (2004) both found that tasks which could be characterised as fault-critical (for example, when participants were describing an object for an interlocutor who had to learn how to use it) resulted in more overspecification. In addition, Arts also cites evidence that locative expressions are frequently used in descriptions and can facilitate the task of a listener in resolving a referent. In the present experiment, the reasoning was that if the +FC condition results in more overspecification, there should be more of this when LOCATION is available (+FC+LOC) than when it isn't (+FC−LOC). The question of interest is therefore whether people would use locative expressions in conjunction with MD attributes. There is a sense in which these attributes are qualitatively different from those related to the position of an object. COLOUR, ORIENTATION and SIZE could be characterised as **inherent visual attributes**, that is, they are dimensions that relate to the 'what' of an object. This applies even to SIZE which, though its value as *large* or *small* requires comparison to other objects, still has an absolute (numeric) value and is a feature of the makeup of an entity. In contrast, LOCATION corresponds to

Condition	Fault-Critical	Location
+FC+LOC	yes	yes
+FC−LOC	yes	no
−FC−LOC	no	no

Table 3.3: Between-groups variables in the experiment

the 'where' of an entity, a property that is external to it. There is the possibility that these two kinds of attributes are dealt with at different points in the process of identifying objects, since in order to analyse the inherent perceptual attributes of an entity, it is necessary to first orient attention to where the entity is. This hypothesis is incporated in computational models of visual attention, such as the foundational Feature Integration Model (Treisman and Gelade, 1980), which is based on the view that simple visual features are computed in parallel, and attention (including spatial orientation) is necessary to focus on an object and bind those features into an object representation (see Itti, 2005, for a recent summary of this model). In this model, then, orienting to the 'where' of an entity is prior to attending to the 'what'.

Since all objects had a unique location in each domain this property is potentially minimally distinguishing, especially if both X- and Y-DIMENSION are used. No two objects had the same numeric values on both dimensions. However, the distinguishing character of locative expressions would depend on whether authors used it in a precise way, for example by giving row and column numbers. This was considered unlikely, since out of 15 grid cells, only a maximum of 8 would be filled by domain objects (at most 2 targets and 6 distractors). Since participants did not see the grid-lines, they would be more likely to use expressions such as *left* or *towards the bottom*, rather than *row 1, column 2*. The two locative attributes were therefore considered 'gradable', in the sense that these expressions represent intervals on the relevant dimension.

3.3 Hypotheses

At the outset of the experiment, three hypotheses were formulated, motivated in part by previous research, and in part by their relevance to GRE.

H1 **Attribute preferences**

1. TYPE and COLOUR will be used in descriptions even when they are unnecessary. ORIENTATION and SIZE will be less frequent as overspecified attributes, with the relative attribute SIZE coming last. Thus, the data should support the following preference order: TYPE >> COLOUR >> ORIENTATION >> SIZE
2. Correspondingly, there will be a greater tendency to *under*specify when SIZE or ORIENTATION are required to identify the referents, as compared to when COLOUR is required.
3. LOCATION will be used whenever possible, and Y-DIMENSION will be used more than X-DIMENSION.

H2 **Effect of communicative setting**

1. There will be more overspecification in +FC conditions.
2. The possibility of rectifying referent identification failures in −FC−LOC will result in higher proportions of underspecified descriptions.

H3 **Effect of cardinality and similarity** Participants will be as likely to overspecify or underspecify descriptions in the Plural conditions as in the Singular, irrespective of the level of Similarity.

The hypotheses about attribute preferences are mostly based on the psycholinguistic literature. As discussed in Chapter 2, the tendency to overspecify descriptions depends on the attribute in question, with COLOUR frequently cited as a preferred attribute (e.g. Pechmann, 1989), while attributes with low codability, such as SIZE, tend to be filtered out unless absolutely required (Belke and Meyer, 2002). Despite SIZE being binary-valued in the domains, using this attribute would still require comparison between objects. ORIENTATION was expected to fall somewhere between these two, given that *left* and *right* values were likely to require mental rotation of the objects. From the point of view of GRE, H1 is relevant in order to determine which version(s) of the IA should be tested against the data. With 4 different attributes (COLOUR, ORIENTATION, SIZE and TYPE), there are 24 possible orders, increasing to 120 possible orderings once the two location attributes (X-DIMENSION and Y-DIMENSION) are included. Clearly, evaluating 120 different IAs would be impractical, and statistically questionable, since it could result in a combinatorial explosion in the number of statistical hypotheses tested, compromising any significant findings (see §4.3.3, p. 111).

A second motivation for H1 is to attempt to replicate previous findings on overspecification in varied communicative settings, as a function of FC, and with plurals. The +FC condition is expected to yield more evidence of overspecification, since the instructions in this condition were worded in a way that emphasised clarity and avoidance of ambiguity. However, the use of LOCATION should interact in interesting ways with this tendency. A replication of the finding by Arts (2004) to the effect that LOCATION is used whenever possible, particularly in the vertical dimension (H1-3), may be accompanied by a decrease in the usage of other properties. Such a result would suggest that, while the formation of a Gestalt in Pechmann's (1989) sense requires processing the visual inherent visual properties of an object, this process may be interrupted if the location of an object (to which an author has already oriented their attention) is sufficient to identify it. H1 therefore addresses Pechmann's *Gestalts Hypothesis* (Pechmann, 1989), which explains overspecification in reference and preference for certain attributes as the result of the way objects are perceived and mentally represented, not as sets of separable dimensions, but as 'more than the sum of their parts', to use the term from Gestalt perception that the hypothesis echoes (e.g. Wertheimer, 1938). In this representation, highly codable attributes are central, but the hypothesis also addresses the case where such attributes are omitted because of perceptual processes that precede the computation of a unified representation of the object using these attributes.

Questions about reference in visual domains have never been addressed for the case of plurals. Whether overspecification is as likely in this case as with singulars remains an open question. If the *Gestalts Hypothesis* is correct, descriptions of sets of objects should also be overspecified

(H3). There are two reasons why this hypothesis is non-trivial. First, it is possible that pluralities incur a higher cognitive cost in searching for distinguishing properties, increasing the likelihood of overspecification. Second, Pechmann's *Gestalts Hypothesis* leaves open two possibilities in the case of plurals: Subjects may either process individual elements of the plurality separately, as it were partitioning the set of referents and constructing a description for each element of the partition. Alternatively, authors may perceive the set as one whole aggregate, and describe the set as a whole. This would presumably depend on the similarity of the target referents. An effect of Similarity would indicate that this is indeed the case, and that algorithms based on partitioning, such as those proposed by Horacek (2004) and van Deemter and Krahmer (2006), are on the right track.

With the exception of research on dialogue (Goodman, 1986; Campana et al., 2002), studies have tended to focus on what extra information experimental participants use in reference, in violation of a strict interpretation of the Gricean Quantity maxim. However, the dependent variables used to test the above hypotheses include *under*specification. Underspecification is interesting for several reasons. First, it may strengthen observations about attribute preferences: if an attribute is easy to process and tends to be included when not required, it should also not be omitted when required. However, a situation in which overspecification and omission conflict is conceivable, and this would complicate the determination of a preference ordering for the IA. Second, while most psycholinguistic experiments reviewed in §2.6.1 (p. 38) relied on relatively small domains, often with one distractor for a single target referent, more complex domains such as the ones used here, where there were 6 distractor objects, may well result in more unsuccessful references. Third, participants in the +LOC condition have the option of not using the three predefined attributes of COLOUR, ORIENTATION and SIZE.

3.4 Participants and procedure

The experiment was run over the Internet, over a period of three months. It was linked from the personal web pages of the participants of the TUNA project, and from two web experiment portals, one at the University of Edinburgh[4] and the other at the University of Zurich [5].

Participants were required to complete the experiment in one sitting; data from people who failed to complete it was not included in the corpus. The study reported here formed part of a larger collaborative effort, which included another domain apart from furniture, consisting of real photographs of people. There were 18 trials in this domain, so that participants completed $(20 + 18 =)$ 38 trials in total. These were administered as unblocked randomised trials for each participant[6].

The experiment consisted of three principal stages. First, participants were asked to give a few details, including a self-report of their fluency in English. They were then randomly assigned to one of the three conditions, and shown the corresponding instructions. In every version, participants were told that the experiment was being conducted to test a language understanding system, which would interpret their descriptions and then remove the target referents from the domain. In

[4] The Language Experiments Portal: `http://www.language-experiments.org/`

[5] The Web Experimental Psychology Lab: `http://www.psychologie.unizh.ch/sowi/Ulf/Lab/WebExpPsyLab.html`.

[6] That is, in the randomisation, furniture/household and people domains were not kept as separate blocks.

actual fact, the system's correct or incorrect responses were predefined, such that the system would respond incorrectly approximately one fourth of the time, removing an arbitrary set of objects[7]. In the instructions, participants were asked to type their descriptions as if interacting remotely with another person. The main part of the instructions, which sought to emphasise *identification* as the principal communicative aim, was as follows:

> In this experiment, we are trying to evaluate the performance of a computer program that understands English. You will be given a task in which you describe and identify objects for the computer. The computer will then try to interpret your description.
> ...
> Each time you do this, click the *submit* button. The program will then try to figure out which objects you mean, and remove them from the screen. It can "see" exactly the same pictures as you ...

After the instructions, trials were administered in a random order determined for each participant. For each trial, the position of objects in cells of the 3×5 grid was randomly determined at runtime. The target referents in the domain were surrounded by a red outline of a width of 2 pixels. In order to ensure that size differences would not be lost due to browser resizing at client-side, the difference between pictures corresponding to the *small* and *large* values of the SIZE attribute was specified as a relative percentage value.

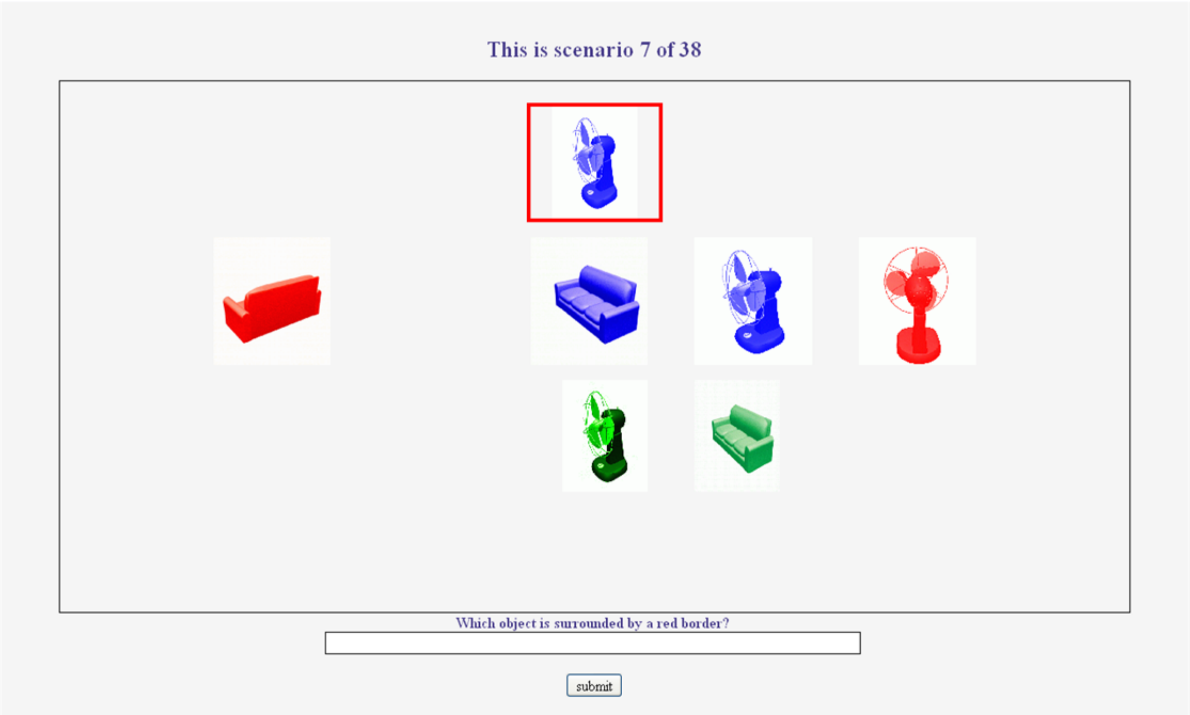
The domain as a whole was surrounded by a black border, beneath which was the question *Which object(s) is/are surrounded by a red border?*. Participants replied by typing a description into a text area. This input phase was followed by a feedback phase: On submission, a function removed the target(s) (or an arbitrary set of objects in case the system made an error) from the domain, after a timeout of $1500ms$. At this point, participants in the −FC−LOC condition were given the opportunity to correct the system's mistakes in case it had identified the wrong targets, by clicking on the right pictures, which were immediately removed from the domain. No such option was given to participants in the +FC conditions. Figure 3.2(a) displays a trial, and Figure 3.2(b) shows the feedback page immediately following the trial, one in which the system had removed the correct targets.

Since this experiment was conducted remotely, some further precautions were taken. Trials could not be repeated (for example, by refreshing the browser and resubmitting a description), and participants could not proceed past a trial without having submitted a description.

The third and final stage of the experiment was a debriefing page in which participants were asked if they had experienced any technical problems during the experiment, and whether they would like their data to be included in the analysis or discarded. Because of the importance of COLOUR on a number of trials, they were also asked whether they had normal colour vision. Finally, they were asked to judge the performance of the system, by rating their agreement to the statement *The system performed well on this task.* on a scale from 1 (*strongly agree*) to 5 (*strongly disagree*). Further comments could be supplied in a text box.

For the compilation of the corpus, a threshold was set, whereby data from a participant would not be included in the corpus if over 20% of descriptions consisted only of TYPE. From a total

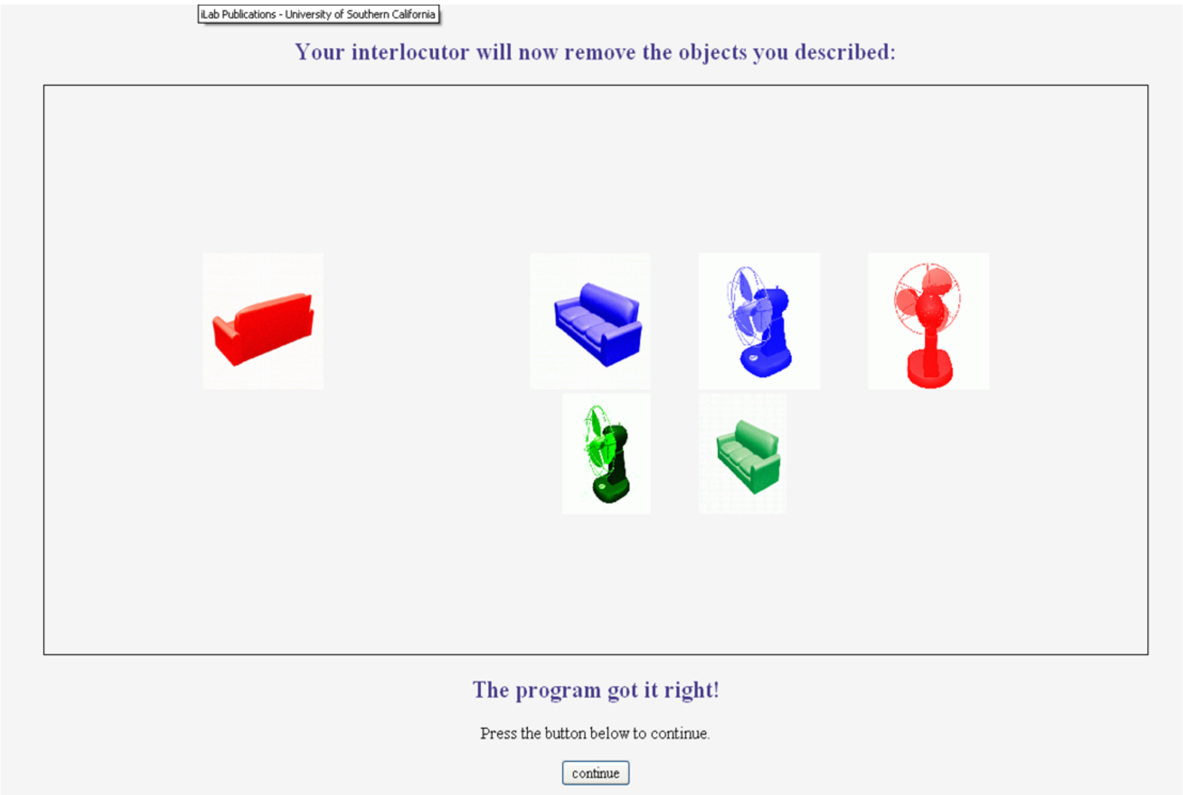
[7] The number of objects removed was always the same as the number of targets.



This is scenario 7 of 38

Which object is surrounded by a red border?

(a) Input phase in an experimental trial



Your interlocutor will now remove the objects you described:

The program got it right!

Press the button below to continue.

(b) Feedback phase in an experimental trial

Figure 3.2: The two phases of an experimental trial: (a) Input phase (b) Feedback phase

of 49 native or fluent speakers of English who completed the experiment, two were omitted by this criterion. A further participant was omitted due to a high degree of syntactic ambiguity in their descriptions, which made annotation impossible, while another explicitly requested during debriefing that their data not be used for analysis. This left 45 participants, 15 in each version, yielding a corpus of $(45 \times 20 =)$ 900 descriptions.

3.4.1 Validity of the experimental method

Since this experiment required participants to interact with a remote system, the perceived success of the system in 'understanding' utterances was deemed crucial to the success of the experiment, since the naturalness of people's descriptions would depend on the extent to which they felt they could really type descriptions 'as if talking to another person'. People's responses to the final part of the debriefing phase, where they rated agreement to the statement *The system performed well on this task*, are a partial indicator of whether this was the case, and also of whether participants had realised that testing the language understanding system was not the true purpose of the experiment.

Of the 5 response categories, ranging from *strongly agree* to *strongly disagree*, 34 (75.5%) individuals selected *agree* or *strongly agree*, while none selected *strongly disagree*. Of the rest, 6 participants selected *neither agree nor disagree*, while only 3 selected *disagree*. This suggests that the setup of the experiment worked reasonably well overall in getting people to interact with a remote system.

3.5 Annotation of the results

Annotation of the data, in XML, was carried out in order to meet the requirements of semantic transparency, enabling the investigation of the hypotheses, and the subsequent evaluation of algorithms. Each corpus description was paired with its domain, which represented all the objects in the trial, identified via an integer ID which corresponded to the picture of the object. Entities in the domain had associated attribute-value information, as well as their location (row and column numbers) in the grid for that specific trial. The annotation scheme devised for the descriptions met the following requirements:

1. The text of a participants' description was left intact.

2. Segments of the text which corresponded to domain properties were tagged in order to make their semantics explicit, while abstracting away from individual differences in lexicalisation and realisation.

3. The annotation also permitted the automatic compilation of a logical form corresponding to the description, to enable direct comparison with the output of a generator.

An annotation scheme that meets these requirements can be used to evaluate content-determination strategies, by exposing algorithms to the same domain as humans, and comparing the results to human data at the semantic level. This form of abstraction away from the NL realisation of human descriptions distinguishes this method from other corpus-based methods, such as that used by Roy (2002). Roy used a corpus of descriptions, elicited against simple domains involving pictures which were generated on the fly, and represented combinations of SHAPE, COLOUR and LOCATION in $2D$ space. He then used machine-learning techniques to map

ID	TYPE	COLOUR	ORIENTATION	SIZE	X-DIMENSION	Y-DIMENSION
84	desk	red	backward	small	3	1
100	sofa	red	backward	small	5	2
20	desk	red	backward	large	1	1
24	desk	red	forward	large	2	3
29	desk	blue	rightward	large	2	4
36	sofa	red	backward	large	4	1
40	sofa	red	forward	large	3	3
45	sofa	blue	rightward	large	3	2

Table 3.4: A domain corresponding to an experimental trial

the human-produced descriptions to the low-level features of the domain objects. For example, COLOUR expressions were paired with the RGB colour values of objects, LOCATION expressions to the $2D$ visual coordinates. The aim was to ground the language directly in perceptual features (see Roy, 2005, for a discussion of the theoretical underpinnings of this approach). By contrast, the purpose of the present annotation scheme was to yield a reliable semantic representation which mediates *between* the linguistic and the perceptual properties of domain objects.

The annotation method, as well as a DTD for the corpus, is documented in van der Sluis et al. (2006). In what follows, I give some examples of relevant features of the annotation scheme, using the domain shown in Table 3.4 as an example. This represents a Plural/Similar trial, with only SIZE in MD, that is, both targets had the same value on the SIZE attribute and this was sufficient to distinguish them. The target referents are shown in boldface. The description given by one author in this domain is shown below.

(3.2) the small red desk and the small sofa facing away

Three kinds of tags were used for the annotation. Segments of a description corresponding to (realisations of) domain properties were enclosed within an `<ATTRIBUTE>` tag. A `<DET>` tag was used for determiners. Sequences of `<ATTRIBUTE>` and `<DET>` tags were enclosed within `<DESCRIPTION>` tags. The text of a participant's description could contain several embedded `<DESCRIPTION>` tags. This was usually the case with coordinate NPs. For example, a description like *the red chair and the red table* would consist of an outer `<DESCRIPTION>` tag with two inner ones, corresponding to the syntactic analysis of the phrase as [[*the red chair*] *and* [*the red table*]]. Further examples of the use of the `<DESCRIPTION>` tag are given below.

`<ATTRIBUTE>` tags had two XML attributes, `name` and `value`, which corresponded to the name of the domain attribute, and the value used by a participant in their description. In addition, `name` took the value `other` when a participant described an aspect of a picture with an attribute that did not clearly have a counterpart in the actual domain, as originally represented. Because of the well-defined nature of the domains, this was only necessary in 39 descriptions (3.2%). A tag's `value` could also be annotated as `other`. For instance, some participants referred to domain objects as *the picture*. This was considered an *other* value of TYPE.

Locative expressions were marked up with a special `<ATTRIBUTE>` tag which could receive numeric values for `x-dimension` and `y-dimension`, apart from the `name` and `value` XML attributes. `<DET>` had a `value` attribute, which indicated the type of determiner used, whether *definite, indefinite, quantificational, numeral* or *other*.

For the example in (3.2), the annotation is shown in (3.3).

(3.3)
```
<DET value=''definite''>the</DET>
 <ATTRIBUTE name=''size'' value=''small''>small</ATTRIBUTE>
 <ATTRIBUTE name=''colour'' value=''red''>red</ATTRIBUTE>
 <ATTRIBUTE name=''type'' value=''desk''>desk</ATTRIBUTE>
and
<DET value=''definite''>the</DET>
 <ATTRIBUTE name=''size'' value=''small''>small</ATTRIBUTE>
 <ATTRIBUTE name=''type'' value=''sofa''>sofa</ATTRIBUTE>
 <ATTRIBUTE name=''orientation'' value=''backward''>
  facing away
 </ATTRIBUTE>
```

The application of `<ATTRIBUTE>` tags involved some interpretation on the part of the annotator, in that it was necessary to determine what domain properties had been expressed by a participant. However, because the domain was well-defined, this proved to be a straightforward exercise in most instances (see §3.5.1). Whenever unclarities arose as to the *value* of an expression corresponding to a domain attribute (say, COLOUR), they could generally be resolved with reference to the domain. For instance, one participant consistently used *purple* for COLOUR, whenever the domain specified *blue* for the target. The `value` here was annotated as *blue*. Similarly, values such as *medium-sized* for an object specified as *large* were annotated with the latter value for SIZE. The reason was that such properties were presumably used because of their contrastive value; comparing algorithms against this data required that the domains on which human and algorithm had produced references be compatible whenever possible. The exception, of course, was when expressions contained attributes that were not specified in the domain at all; these were tagged using `name=''other''`.

LOCATION information was the most complex to annotate. The `value` attribute of `<ATTRIBUTE name=''location''>` took as a value one of `top`, `bottom`, `left`, `right`, `middle` or `other`. The use of these values reflected the degree to which an expression specified properties like *top* and *right* explicitly. Any locative expression that corresponded to the horizontal dimension (*left*, *right* or *middle* when this referred to the middle of a row) also had the numeric `x-dimension` attribute. Similarly, expressions corresponding to the vertical dimension had the numeric `y-dimension` attribute. All other expressions had both. In case a locative expression described the position of an object relative to another, an additional `rel` attribute specified the ID of the relatum. Parts of a locative expression corresponding to different dimensions were annotated separately. These different cases are exemplified in (3.4–3.6), in which the annotation for locative expressions (shown in boldface) is shown. These are hypothetical references to entity 84 in Table 3.4.

(3.4) the **topmost** desk
```
<ATTRIBUTE name=''location'' value=''top'' y-dimension=''1''>
 topmost
</ATTRIBUTE>
```

(3.5) the desk **at the top middle**

```
<ATTRIBUTE name=''location'' value=''top'' y-dimension=''1''>
 top
</ATTRIBUTE>
<ATTRIBUTE name=''location'' value=''middle'' x-dimension=''3''>
 middle
</ATTRIBUTE>
```

(3.6) the desk **above the blue sofa**

```
<ATTRIBUTE name=''location'' value=''top'' y-dimension=''1''
rel=''45''>
 above the blue sofa
</ATTRIBUTE>
```

Note that example (3.4) only contains a `y-dimension` specification, in contrast to (3.5). This is because the former clearly specifies only the vertical position of the target. The `rel` attribute in relational expressions such as (3.6), could also generally be resolved via the domain. Here, the blue sofa which is just below the target is object 45.

Every expression in the corpus was enclosed in an outer `<DESCRIPTION>` tag, with further embedded `<DESCRIPTION>` tags in case the expression was a coordinate NP. For example, (3.2) receives the following annotation at this level:

(3.7)

```
<DESCRIPTION NUM=''PLURAL''>
 <DESCRIPTION NUM=''PLURAL''>
  <DESCRIPTION NUM=''SINGULAR''>
   the small red desk
  </DESCRIPTION>
 and
  <DESCRIPTION NUM=''SINGULAR''>
   the small sofa
  </DESCRIPTION>
 </DESCRIPTION>
 facing away
</DESCRIPTION>
```

One of the functions of the `<DESCRIPTION>` tag was to make the structure of an expression transparent enough for a system to automatically compile the corresponding logical form from the annotation. In example (3.7), the whole is enclosed in a plural description tag, indicated by the `num` attribute. Each coordinate NP is enclosed in a singular description tag. The expression corresponding to the ORIENTATION attribute, *facing away*, is syntactically ambiguous: it can be interpreted as modifying either both coordinate NPs, or only the second. In cases of modifier attachment ambiguity, the strategy adopted was to take the largest possible segment of an expression as the attachment site, given the domain information. Here, since both targets have the same value for ORIENTATION, *facing away* modifies both. This is made explicit by using an embedded `<DESCRIPTION num=''plural''>` tag, which encloses the coordinate NP. This inner `plural` description is modified by the `<ATTRIBUTE name=''orientation''>` tag, within the outer `<DESCRIPTION>`.

	syntactic rule	**semantic rule**
1.	$D_{sg} \rightarrow A_1, \ldots, A_n$	$[\![\ D_{sg}\ ]\!] = A_1 \wedge \ldots \wedge A_n$
2.	$D_{pl} \rightarrow A_1, \ldots, A_n$	$[\![\ D_{pl}\ ]\!] = A_1 \wedge \ldots \wedge A_n$
3.	$D_{pl} \rightarrow D_1, \ldots, D_n$	$[\![\ D_{pl}\ ]\!] = [\![\ D_1\ ]\!] \vee \ldots \vee [\![\ D_n\ ]\!]$
4.	$D_{pl} \rightarrow D, A$	$[\![\ D_{pl}\ ]\!] = [\![\ D\ ]\!] \wedge A$

Figure 3.3: Rules for the interpretation of descriptions using the XML data

On the basis of the simple syntactic markup, a logical form could be derived compositionally. The derivation of a logical form is achieved by the recursive application of the semantic rules shown in Figure 3.3. The left hand side of the figure shows syntactic rules, in a context-free grammar format. Rules 1 and 2 stipulate that a singular or plural <DESCRIPTION> tag (denoted D_{sg} and D_{pl}) could have any number of <ATTRIBUTE> tags (A) as children. The corresponding semantic form is a conjunction. A plural description can also be composed of several embedded descriptions (rule 3). Description nodes which are siblings in the XML tree are disjoined. On the other hand, a description whose sibling is an attribute node is conjoined to the semantic representation of that node (rule 4). Using these rules, the example description in (3.7) yields the logical form in 3.8).

(3.8)
$$\begin{array}{c}\big[\big(\langle \text{SIZE} : \textit{small}\rangle \wedge \langle \text{COLOUR} : \textit{red}\rangle \wedge \langle \text{TYPE} : \textit{desk}\rangle\big) \\ \vee \\ \big(\langle\langle \text{SIZE} : \textit{small}\rangle \wedge \langle \text{TYPE} : \textit{sofa}\rangle\big)\big] \\ \wedge \\ \langle \text{ORIENTATION} : \textit{backward}\rangle\end{array}$$

3.5.1 Annotation procedure and inter-annotator reliability

The data was divided into two halves and each half was annotated by one person[8]. Each annotator's results were then validated by the other, and corrections made where disagreements arose regarding ambiguity.

The reliability of the annotation scheme was subsequently evaluated in a study involving two independent annotators, both postgraduate students with an interest in NLG. The annotators (hereafter A and B) were given the manual used for the annotation of the corpus (van der Sluis et al., 2006), and given a brief introduction to the annotation task, using 8 example descriptions. No further training was given. The data used for the study was a stratified random sample of 270 descriptions, 2 from each of the Singular, Plural Similar and Plural Dissimilar conditions, from each author in the corpus.

To compare A and B's annotations to those in the corpus, each annotated description was compiled into a logical form, using the rules in Figure 3.3. The resulting logical form was compared to the one obtained from the corpus annotation, using the Dice coefficient of similarity. Let D_1 and D_2 be two descriptions, and $att(D)$ be the attributes in any description D. The coefficient, which ranges between 0 (no agreement) and 1 (perfect agreement) is calculated as in (3.9). Because descriptions could contain more than one instance of an attribute (e.g. Figure 3.8 contains

[8] The annotators were the present author, and another member of the TUNA project.

	Annotator A	**Annotator** B
mean	0.93	0.92
mode	1	1
PRP	74.4%	73%

Table 3.5: Mean and modal inter-annotator agreement scores

two instances of SIZE), the sets of attributes for this comparison were represented as multisets.

$$dice(D_1, D_2) = \frac{2 \times |att(D_1) \cap att(D_2)|}{|att(D_1)| + |att(D_2)|} \tag{3.9}$$

The reason for using this coefficient, as opposed to one of the more familiar agreement statistics, such as Kappa (κ), is that the latter requires a collection of well-defined discrete events which are placed into predefined categories. In the present case, the categories (the attributes in the annotation) were indeed predefined; however, the events that were classified consisted of bits of language. The number of such events in any corpus description depended on an annotator's intuitions given the annotation instructions. In other words, the corpus data did not permit an *a priori* segmentation of the descriptions, and a subsequent classification of the segments, since segmentation would in itself be tantamount to a judgement of the 'attribute-hood' of a particular sequence of words.

Both annotators showed a high agreement with the TUNA annotators, as indicated by the mean and modal (most frequent) agreement scores shown in Table 3.5. The table also shows the *perfect recall percentage* (PRP), that is the proportion of times an annotator agreed perfectly with the corpus annotation (this happened to coincide with the modal score in both cases). There was perfect agreement on attribute annotation over 70% of the time, with a mean that was close to the perfect match in each case. In addition, both annotators evinced a substantial agreement among themselves (mean $= 0.89$, mode $= 1$ (71.1%)). The results therefore suggest that the annotation scheme used is replicable to a high degree, and that independent annotators are likely to produce similar semantic markup.

3.6 Data analysis

This section addresses the hypotheses outlined in §3.3. The outcome of this analysis will then inform the evaluation experiment reported in the next Chapter. The data analysis is divided into three main parts, according to the three main hypotheses stated at the outset.

Hypothesis H1 was related to attribute preferences. To investigate these, I will focus on each of the three main attributes along which objects were defined in the corpus domains, namely COLOUR, ORIENTATION and SIZE, as well as LOCATION where relevant. The main dependent variables here are the proportion of descriptions that *include an attribute when it is not required* (hereafter *attribute overspecification*) and the proportion of descriptions that omit an attribute *when it is required*, that is, when it forms part of MD (*attribute omission*). Proportions of overspecified uses of the three attributes were calculated by identifying those instances where participants used such attributes in domains where they were not required (i.e. were not part of MD). For

COLOUR and ORIENTATION, there were exactly 8 such trials, with 9 for SIZE (cf. §3.2.2). Similarly, proportions of times that participants underspecified by omitting either of these attributes were obtained by considering descriptions that did not include an attribute where it was required (12 possible trials for each of COLOUR and ORIENTATION; 11 for SIZE).

Attribute omissions need not result in underspecified descriptions in the strict sense, since the use of LOCATION can make up for the loss of information incurred when an attribute is omitted. However, omission and overspecified usage of attributes are useful indicators of the extent to which an attribute is preferred or dispreferred.

The second and third hypotheses were related to over- and underspecification in different levels of the Condition and Cardinality/Similarity factors. To investigate H2 and H3, I will use proportions of *overspecified and underspecified descriptions*. These categories overlap with those in which there is overspecified usage or omission of MD attributes, but are not quite the same, because they take LOCATION into account. Explicit definitions of the categories is provided below. Another indicator of informativeness of descriptions in these conditions is *description length*, defined as the total number of attributes in a description.

Unless otherwise stated, the statistical analyses carried out on proportions of response type, reported using participants and items as sources of variance[9]. Categorical data such as this tends not to permit the assumptions that underly parametric statistics. In particular, zero values are expected to be frequent (when there are no responses in a given category), and there is a high dependency of variance on the mean[10]. Thus, non-parametric statistics are used for response proportions. For overall comparisons, I report the results of Friedman Analysis of Variance, and Kruskall-Wallis tests for between-groups analyses (both denoted χ^2_1 for the by-subjects analysis, and χ^2_2 for the by-items analysis). Where significant main effects are found, I use Signed Ranks tests, again by subjects (Z_1) and items (Z_2), for pairwise comparisons to further investigate the nature of the effects. In the case of description length, since this is a scale variable and is normal, I report the results of parametric tests to compare means.

On those occasions when the corpus data indicates patterns which, though not directly predicted in H1 to H3, are relevant to them, I will report percentages of the relevant occurrences, and use simple χ^2 tests on frequencies in the corpus to check their reliability.

3.6.1 Overview of the data

The XML corpus data was post-processed to determine which descriptions were overspecified and underspecified in relation to the domain. To determine whether a description is underspecified, it is necessary to take into account (a) whether it includes the attributes required by MD; (b) whether it includes a locative expression, that is, an expression consisting of X-DIMENSION, Y-DIMENSION, or both. The definitions of overspecification and underspecification used are given below.

1. A description is *underspecified* if it does not include LOCATION and it omits required MD attributes. For example (3.10) is underspecified, because the domain here was such that both SIZE and COLOUR were required to distinguish the targets (MD = COLOUR+SIZE), but

[9] A by-subjects analysis is carried out by averaging the dependent variable of interest per participant, for each level of the independent variables. Similarly, the by-items analysis involves averaging over items. In this case, items are the 20 domains in the corpus.

[10] Samples from a normally distributed population will have differing means, but constant variance. This is the primary assumption in parametric statistics.

the description does not contain SIZE information, and there is no locative expression that makes up for the loss of information. In such cases, once MD attributes are omitted, using other inherent attributes without using LOCATION will still make the description ambiguous (i.e. non-distinguishing).

(3.10) a green fan, a green chair

2. A description is *overspecified* if either of the following is true:

   (a) The description does not omit any MD attributes, but includes LOCATION or extra MD attributes;

   (b) The description omits some MD attributes, but includes both LOCATION and extra MD attributes.

   The rationale behind this definition was that LOCATION could often distinguish a referent uniquely; therefore, any description that included this attribute over and above the minimally required inherent visual properties, or one that included this attribute together with other inherent visual properties, contained more information than strictly necessary to identify the referent(s). Example (3.11) is a case of overspecification, produced in a domain in which ORIENTATION alone was required to identify the referent. The extra information is COLOUR.

   (3.11) the red chair shown with the seat on the right

3. A description is *well-specified* if it is neither overspecified nor underspecified. Description (3.11c) is well-specified, because it was produced in a domain where colour alone sufficed to distinguish the target referent, and it contains no further information.

   (3.12) grey desk

In identifying over- and underspecified descriptions, TYPE was never taken into account as it was assumed that this was required on independent grounds (see §3.8 for discussion). On the other hand, the *length* of a description was defined as the total number of attributes used in it, including TYPE.

In determining whether a description is overspecified, underspecified or neither, LOCATION complicates the picture somewhat. As the definitions show, the stance taken here is conservative, based on the assumption that if a locative was used, then this counted as additional information that could make up for the omission of MD attributes. Of course, this usually depended on whether the locative expression itself was adequate to identify the target referents, either on its own or in conjunction with other properties in the description. This is difficult to determine objectively. One strategy would be to look at the corpus annotations, and determine whether a locative expression used X-DIMENSION, Y-DIMENSION, or both, and whether either of these was sufficient to identify the targets. This would involve checking whether the numeric value of the target on the dimensions used was unique for the target. For example, the referent might be the only entity in column 4 in a domain. Hence, $\langle$X-DIMENSION $=$ *4*$\rangle$ is sufficient to distinguish it. On this basis, if a person added Y-DIMENSION to the description, apart from X-DIMENSION, this would make

	Well-specified	**Overspecified**	**Underspecified**
+FC+LOC	66	29.7	4.3
+FC−LOC	51.3	39.7	9
−FC−LOC	46.7	33.3	20
overall	54.7	34.2	11.1

Table 3.6: Percentage overspecified, underspecified and well-specified descriptions

the description overspecified. However, this would not be true to the data. The annotation with numeric position attributes does not take into account the fact that a large number of locative expressions were 'vague'. This is clear from the examples below, where the segments in boldface suggest that LOCATION was seen as a gradable attribute by at least some individuals.

(3.13) (a) in the middle, **towards the left**

(b) on top, **slightly towards the right**

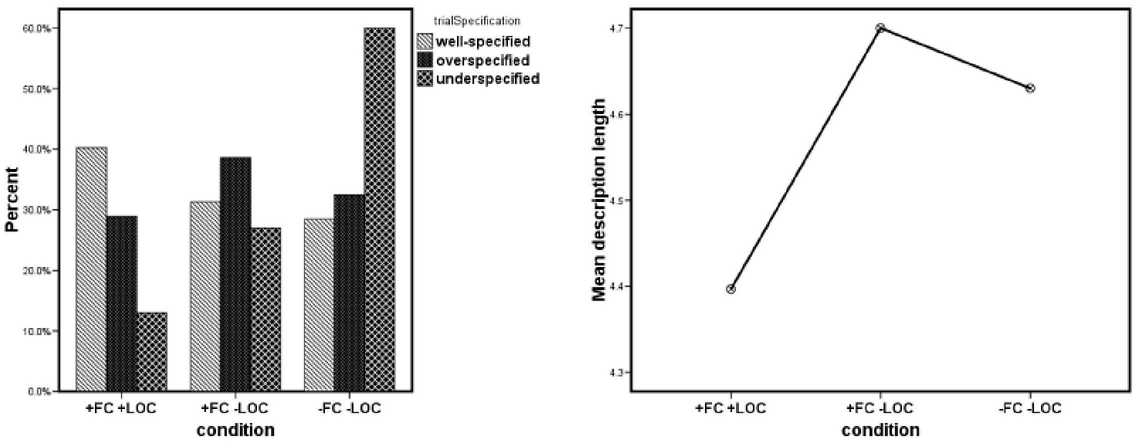
Because of the apparently gradable nature of locative expressions, I decided to avoid simple decisions as to whether such expressions were fully distinguishing or not. I return to the question of locatives in §3.8.1. Here, it is worth noting that LOCATION was used with very high frequency. In the +FC+LOC condition, it was used consistently by over half (50.4%) the participants. It was also used by 12 participants out of the 30 in the −LOC conditions, 8 of whom were in −FC−LOC. These participants used it, usually in conjunction with other properties, on an average of 59% of their descriptions, despite being given instructions to the contrary.

The proportions of overspecified, underspecified and well-specified descriptions in the entire corpus are shown in Table 3.6, which also displays proportions per Condition. As shown in the table, there were relatively few underspecified descriptions in the corpus overall, and the majority of these occurred in the −FC−LOC condition. The raw figures also suggest that the greatest proportion of overspecified descriptions was to be found in the +FC−LOC condition, whereas the other fault-critical condition (where participants had the option to use LOCATION), evinces less overspecification. Thus, there are apparent differences among conditions which seem to go in the direction predicted by Hypothesis H2. Before going into the question of attribute preferences in relation to H1, then, I first look at the effects of Condition and fault-criticalness on people's descriptions.

3.7 The effects of condition and fault-criticalness

The proportions of descriptions given in Table 3.6 are shown graphically in Figure 3.4(a). Figure 3.4(b) displays the mean length of descriptions per condition. Although the Figure shows a difference in length between conditions, suggesting that people tended to produce lengthier descriptions when they could not use LOCATION, the difference is not dramatic. A one-way ANOVA using Condition as independent variable showed that the differences in length were not reliable ($F_1(2, 44) = .5$, ns; $F_2(2, 59) = .2$, ns).

The picture was somewhat different with the proportions of well-specified, overspecified, and underspecified descriptions. The effect of Condition on proportions of well-specified descriptions



(a) Specification of descriptions

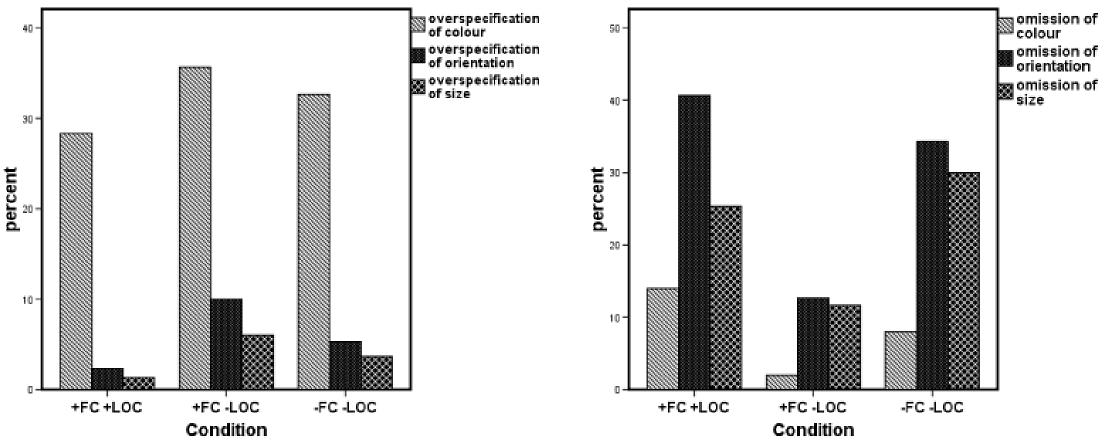
(b) Mean description length

Figure 3.4: Properties of descriptions as a function of Condition

approached significance by subjects ($\chi_1^2 = 5.538$, $p = .06$), and was significant by items ($\chi_1^2 = 7.385$, $p = .03$). The same pattern was observed with proportions of underspecified responses ($\chi_1^2 = 4.855$, $p = .08$; $\chi_2^2 = 20.6$, $p < .001$). There was no difference between conditions on proportions of overspecified responses. Pairwise comparisons revealed that the main effect of Condition on well-specified responses was due solely to a difference between +FC+LOC and −FC−LOC ($Z_1 = 2.210$, $p = .03$; $Z_2 = 2.825$, $p = .004$). Thus, the use of LOCATION in the +FC+LOC condition resulted in a reliable increase in the proportions of well-specified responses as defined above. Further investigation of the effect of Condition on underspecified descriptions revealed reliably more of these in −FC−LOC compared to −FC+LOC, though only by subjects ($Z_1 = 2.071$, $p = .05$; $Z_2 = 1.161$, ns). Similarly, authors in −FC−LOC were more likely to underspecify than in +FC+LOC. This difference approached significance by subjects ($Z_1 = 1.806$, $p = .07$), but was highly reliable by items ($Z_2 = 4.464$, $p < .001$). On proportions of underspecified descriptions, +FC−LOC also differed from the −FC condition, though only by items ($Z_2 = 3.046, p = .002$).

The results suggest that being in a fault-critical situation, and/or having recourse to LOCATION to describe objects, did not result in longer descriptions. A more fine-grained view of what constitutes an overspecified and/or an underspecified description shows some effects of Condition. In particular, the use of LOCATION in the fault-critical condition resulted in more *well-specified* descriptions – those which were neither overspecified nor underspecified – and also in less likelihood of underspecification. In contrast, though the frequency with which people overspecified in +FC−LOC was greater, it failed to reach significance; however, there was also a tendency to underspecify less in this condition, compared to the non-fault-critical one. The results are not uniform, in that there are differences between the by-subjects and the by-items analysis. This implies that although there was some consistency in the type of description produced on a given domain (or 'item'), there was high variability among authors in their choice of referring expression.

The results reported above must be interpreted with the caveats pointed out in §3.6.1 with



(a) Overspecified use of attributes (b) Omission of attributes when required

Figure 3.5: Overspecification and omission of MD attributes

	Overspecification		**Omission**	
	Z_1	Z_2	Z_1	Z_2
+FC+LOC vs. +FC−LOC	2.191^{**}	3.311^{*}	2.135^{**}	3.832^{*}
+FC+LOC vs −FC−LOC	1.462	2.974^{*}	$.028$	$.343$
+FC−LOC vs −FC−LOC	$.920$	2.423^{*}	2.617^{*}	3.931^{*}

Table 3.7: Pairwise comparisons of over- and underspecified responses across versions. ($^{*}p \leq .004$; $^{**}p \leq .05$)

regard to the LOCATION attribute. A slightly different view is afforded by the data on when people omitted MD attributes from their descriptions, versus when they overspecified using these attributes. Figure 3.5 displays proportions of descriptions in which authors omitted SIZE, COLOUR, and ORIENTATION when it was part of MD, or included these attributes when they were not part of MD.

The figure shows a marked difference between the three conditions in the extent to which certain attributes were omitted or included. Authors overspecified least of all in the +FC+LOC condition. This is also the condition where they showed the greatest tendency to omit attributes such as ORIENTATION and COLOUR. Taken together with the high proportion of well-specified descriptions in this condition (see Figure 3.4), this means that people tended to prefer using LOCATION. In contrast, people tended to include extra information much more in the +FC−LOC condition, again confirming the pattern in Figure 3.4, though this turned out not to be significant when LOCATION was factored in. Omission of MD attributes in this condition was correspondingly low.

Condition exerted a significant influence on likelihood of inclusion of extra information provided by the inherent visual properties of objects ($\chi^2_1 = 6.898, p = .03$; $\chi^2_2 = 18.655, p < .001$), as well as on likelihood of omission of such properties ($\chi^2_1 = 9.186$, $p = .01$; $\chi^2_2 = 30.427$, $p < .001$). The pairwise comparisons between conditions are displayed in Table 3.7.

The most reliable differences are those between +FC−LOC and the other two conditions.

Hence, a fault-critical communicative setting without the use of LOCATION resulted in much higher likelihood of inclusion of visual properties, and lower probability of omitting such properties. The relatively weak contrast between +FC+LOC and −FC−LOC (only reliably different on overspecification by items), is partially due to the high proportion of uses of LOCATION in the latter condition. However, the contrast between +FC+LOC and +FC−LOC is of most interest, since the only difference between these two conditions was in whether the authors communicated with an interlocutor who had access to the same locative information.

The results on overspecified usage and omission of MD attributes, together with those on proportions of over- and underspecified descriptions, suggest that the use of LOCATION played a major role in the difference between conditions. When this is factored in as part of the calculation, the results primarily show a difference between proportions of well-specified and underspecified descriptions. With respect to the latter, fault-criticalness is clearly playing a role: authors in +FC−LOC underspecified significantly less than authors in −FC−LOC, omitting MD attributes fewer times, while people in +FC+LOC were more likely to have well-specified descriptions compared to authors in −FC−LOC. However, an important reason for this – as the data on omission of MD attributes shows – is that the use of LOCATION allowed people to identify objects without reference to their inherent visual properties. This might suggest a dominance of the the 'where' of an object over the 'what'. Yet Figure 3.5 also shows some clear preferences for certain attributes over others. It is to these preferences that the analysis now turns.

3.8 Attribute preferences

This section explores the differences between different attributes in more depth. I begin with an analysis of *inherent visual attributes* – those accounted for in MD for a domain – and then move on to a comparison of these attributes to LOCATION.

The definitions of over- and underspecification in §3.6.1 did not take TYPE into account. In part this was because domains were set up in such a way that this attribute would not have any contrastive value. The other reason was that people were expected to use this attribute irrespective of whether it was useful in removing distractors. Participants only omitted TYPE in 6.5% of cases overall. The majority used the predicted values of this attribute (see Table 3.1), with descriptions such as (3.14a–b). Cases where participants used `other` values, exemplified in (3.14c–d), were in the minority (7.8%).

(3.14) (a) red **desk** front-facing and red **chair** front-facing

(b) the only blue **fan** in the middle row and the only blue **couch** on the bottom row

(c) middle **picture** on row 1

(d) two blue **pictures**

The use of TYPE, despite its lack of discriminatory value on any trial, is unsurprising: several authors, including Dale and Reiter (1995), have highlighted the special status of this attribute. Its use conforms to Pechmann's (1989) hypothesis that people perceive objects as gestalts, and that TYPE forms a crucial part of the conceptual representation of entities. The fact that people

	Overspecification		Omission	
	Z_1	Z_2	Z_1	Z_2
SIZE VS. COLOUR	5.695^*	2.555^{***}	4.934^*	2.937^{**}
ORIENTATION VS. COLOUR	5.603^*	2.536^{***}	4.973^*	3.070^{**}
ORIENTATION VS. SIZE	1.881	2.388^{***}	2.933^{**}	2.524^{***}

Table 3.8: Pairwise comparisons of overspecification and omission of different attributes ($^*p \leq .001$; $^{**}p \leq .003$; $^{***}p \leq .05$)

preferred the basic-level values in Table 3.1 also conforms to well-established models of conceptual processing and lexical choice (Rosch et al., 1976; Cruse, 1977; Murphy, 2002) and supports the addition of the *findBestValue*(A) function in the Dale and Reiter model (cf. §2.7, p. 47). There is also a syntactic interpretation that goes hand-in-hand with the gestalts hypothesis: psycholinguistically-motivated computational models of incremental syntactic production, such as Kempen and Hoenkamp (1987), view phrase construction as head-driven. Thus, noun phrases are constructed by first inserting a head noun to which modifiers and determiners are attached. The perception of objects as conceptual gestalts would facilitate this process if the representation of entities is centred around the conceptual category to which they belong, and TYPE information is realised as the head noun of the NP.

The rest of this section focuses on the differences between COLOUR, ORIENTATION, and SIZE, focusing on the descriptions shown in Figure 3.5. Figure 3.5(a) suggests that H1 is on the right track, and that the preference order hypothesised in §3.3 is correct, with more overspecified usage of COLOUR overall, and more for ORIENTATION than SIZE. The likelihood of overspecification of different attributes differed reliably by subjects ($\chi^2_1 = 69.853, p < .001$), and by items ($\chi^2_2 = 15.548, p = .001$). The same was true of omission ($\chi^2_1 = 45.597, p < .001$; $\chi^2_2 = 18.558$, $p < .001$).

Pairwise comparisons of the likelihood of inclusion or omission of attributes are shown in Table 3.8. As expected, there were significantly fewer overspecified uses of ORIENTATION and SIZE compared to COLOUR. However, the likelihood of overspecification of ORIENTATION was not reliably higher than SIZE by subjects. Although this could be due to people being as likely to overspecify using SIZE as ORIENTATION, the omission figures suggest otherwise.

As indicated by Figure 3.5(b), there was a markedly lower trend for omission of COLOUR when it was part of MD. As the right-hand panel of Table 3.8 shows, all pairwise differences involving this attribute were significant. The surprising difference is that between ORIENTATION and SIZE: the chart shows the opposite trend to that in Figure 3.5(a), with the proportion of underspecified responses appearing higher for ORIENTATION than for SIZE in all three conditions.

There is therefore some indication that preference for some attributes is not an all-or-none affair. One possibility is that the usage of ORIENTATION depended on its *value*. As hinted in §3.2.2, *left* or *right* values may be more costly to use, because they require mental rotation, which is cognitively more effortful. This is difficult to ascertain from the corpus data, though there was a slightly higher tendency to omit the attribute in the *left-right* cases. 34% of the descriptions in which ORIENTATION was omitted had at least one target facing *right* or *left*, compared to 32%

involving at least one target facing *front* or *back*. Another possibility is that ORIENTATION was omitted more often when LOCATION was used than was SIZE. This is addressed in the next section. Whatever the reasons for the discrepancies, the data does suggest that even in a simple domain with only three attributes, determining a preference order is not a clear-cut affair. One attribute, namely COLOUR, stands out as highly preferred, but the other two seem to be unordered with respect to each other because the overspecification and omission data conflict.

3.8.1 The role of location

As noted in §3.6.1, there is an apparent preference for the 'where' of an object, with a number of participants using LOCATION even when they were asked not to do so. This also had an impact on the comparison between different conditions, where the effect of a fault-critical communicative task was strongly impacted by the use of this attribute. Some authors, notably Arts (2004), have found that LOCATION is an easy option for people to process in referential tasks. Arts's experiment on listeners showed that referring expressions containing locatives, together with other information, lead to shorter identification latencies. More specifically, locatives helped if they involved both vertical and horizontal axes (both X- and Y-DIMENSION in the current terminology), or only the vertical (Y-DIMENSION). Arts's explanation was that locatives helped to orient attention in a physical context, facilitating object identification, and her results suggest that the top-bottom dimension has some perceptual and/or conceptual primacy.

At first sight, these results contrast with some recent proposals in the GRE literature, for example by Kelleher and Kruijff (2006), whose version of the IA places LOCATION last in the preference order (cf. §2.7.3, p. 54). However, these authors were focusing primarily on *relative* proximity, where locative expressions are n-ary. These may be cognitively costly because they require the processing of properties of relata.

The corpus under discussion only included 16 relational locatives (2.7% of the descriptions which included LOCATION), which implies that relative location is indeed more effortful than absolute location. This ties in with some previous work, involving a corpus of references to spatial groups (e.g. *the group of light bulbs at the bottom right*), in which LOCATION was the only distinguishing feature for a set (Gatt, 2006a). Here, there was a greater use of relational locatives when a target group of referents was in close spatial proximity to some other domain object, compared to when it was spatially isolated. In such cases, authors regularly referred to a nearby distractor to identify the target group. However, though distractor proximity increased the likelihood of relations, they were still in the minority in this condition. This finding is also echoed by Viethen and Dale (2006), who collected a small corpus of references to objects in which location was a possible attribute, and reported an incidence of only 13% of relational descriptions overall.

In the remainder of this section, I focus on absolute locatives. There are two questions of interest. The first is the extent to which the use of LOCATION results in a reduction in the use of other attributes (the 'inherent visual properties' of the object). This would give further indications as to the relative preference of different attributes. The second question is related to whether there is indeed a preference for the Y-DIMENSION attribute. If so, then this would confirm Arts's findings, and would show that the preference holds for speakers, as well as hearers.

Some examples of locative expressions from the corpus are shown in (3.15). As Figure 3.6 shows, the overall preference among participants was to use both dimensions, with descriptions

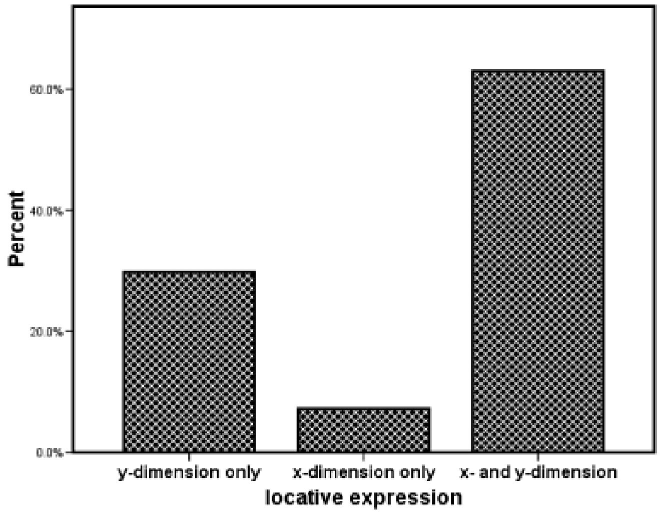


Figure 3.6: Types of locative expressions

such as (3.15a). Among those cases involving only one or the other dimension, there was also a clear preference for Y-DIMENSION, that is, authors preferred to produce descriptions like (3.15b) over (3.15c).

(3.15) (a) right bottom chair

(b) red chair at the top

(c) small green desk on the right

These proportions of uses of the three kinds of locatives in the figure differed significantly ($\chi_1^2 = 14.538$, $p = .001$; $\chi_2^2 = 38.079$, $p < .001$). A comparison of proportions of descriptions containing only Y-DIMENSION and X-DIMENSION showed that the apparent preference for the former in the corpus is reliable ($Z_1 = -2.657$, $p = .008$; $Z_2 = -3.649$, $p < .001$). Overall, it would seem that the use of the horizontal dimension depends on the use of the vertical, that is, if X-DIMENSION is used, it will tend to be used in conjunction with Y. The preference for the use of the vertical dimension is probably due to a combination of two factors. As explained in §3.2.2, objects in the referential domains were placed in a sparse 5 (column) ×3 (row) matrix. Since there were only three rows, determining the vertical position of an entity was easier than determining its horizontal position, especially since participants could not see the grid-lines. Secondly, the preference for vertical dimension may be due to the primacy of the top-bottom continuum over the left-right continuum, as suggested by Arts (2004).

The next question is the extent to which the use of LOCATION resulted in the omission of other attributes. It should be emphasised that here, the focus is *not* on overspecification or underspecification, but on whether a locative expression was more or less likely to be accompanied by other attributes. For this part of the analysis, I focus on whether a description contained COLOUR, ORIENTATION and/or SIZE, as a function of whether it also contained LOCATION. Usage and omission figures are shown in Table 3.9.

	+LOC		−LOC	
	Used	Not used	Used	Not used
COLOUR	62.7	37.3	96	4
SIZE	13.5	86.5	48.9	51
ORIENTATION	6.3	93.7	53.4	46.6

Table 3.9: % Use of MD attributes as a function of the presence of LOCATION

The pattern of use reflects the patterns reported earlier, with COLOUR being used most frequently irrespective of whether LOCATION was included in a description. However, there is once again a discrepancy in the figures for SIZE and ORIENTATION: the latter tended to be used more often compared to SIZE when descriptions did not contain LOCATION; however, SIZE was used more often in those that contained a locative expression.

A more general trend can also be observed in the table: when LOCATION is included in a description, the usage of all other attributes drops. Because the design of the experiment restricted the use of LOCATION to the +FC+LOC condition, I contrasted frequencies of usage of the three attributes in descriptions which contained LOCATION, versus those that didn't, within this condition only, using χ^2 tests on frequencies. Use of LOCATION had a reliable influence on usage of COLOUR ($\chi^2 = 51.955$, $p < .001$), SIZE ($\chi^2 = 37.792$, $p < .001$) and ORIENTATION ($\chi^2 = 77.629, p < .001$).

The results complement the finding of §3.7, where the +LOC condition displayed a greater proportion of well-specified descriptions, among which were those that included LOCATION and omitted some of the visual properties of an object. However, the difference between attributes is noteworthy: use of COLOUR drops by approximately 34%. A much steeper drop is observable for ORIENTATION (47%), while SIZE drops by approximately the same amount, but was far less used in the −LOC descriptions. This data makes the picture about attribute preferences more complete. The patterns can be summarised as follows:

1. When omission of MD attributes and overspecified usage of these inherent visual properties are considered, the preferences suggested by the data are

   COLOUR >> $\{$ ORIENTATION, SIZE $\}$

   with no clear-cut difference between the last two due to a possible preference for specific values of the former.

2. When LOCATION is factored in, there is a clear preference for using the vertical over the horizontal dimension, with the usage of the latter exhibiting a dependency on the former.

3. LOCATION reduces the use of inherent visual properties, but the reduction is smaller for COLOUR which, at 62%, remains very likely to be used.

4. Despite its being overspecified more, ORIENTATION is also omitted more, especially when a description contains a locative expression.

The fact that people focus less on the inherent visual properties of an object when they describe its position could be interpreted in terms of the two processes involved in identifying the properties of an object, one related to its location (the 'where') and another related to its inherent properties (the 'what'). One reason why the former often results in the omission of the latter is that position is often sufficient to identify the referent. Another reason, also suggested by Arts (2004), is that orienting attention to the physical location of an entity is required prior to the recognition and identification process, as predicted by computational models of attention. Thus, unless the 'where' system is suppressed (as when one's interlocutor does not share knowledge of the location of objects), it might have primacy.

The idea that there are two processes at work in identifying objects is supported by the observation that descriptions containing locative expressions were more likely to omit TYPE. Of the descriptions that contained a locative expression, 10% omitted TYPE, compared to only 3.4% of descriptions that did not use LOCATION. Though the difference is not dramatic, a χ^2 test on response frequencies, comparing those descriptions containing or omitting TYPE, as a function of whether they included locative expressions, showed that LOCATION exerted a significant influence on whether to include it ($\chi^2 = 16.423$, $p = .001$). Thus, participants were more likely to use descriptions such as (3.16a) when they included the position of objects than when they didn't. Moreover, though TYPE was used frequently overall, descriptions were more likely to conform to the pattern exemplified in (3.16b), when there was no locative.

(3.16) (a) bottom left, and middle right

(b) smallest red couch and desk

So far, attribute preferences have been interpreted with reference to Pechmann's (1989) Gestalts Hypothesis, which holds that some properties are central to the conceptual representation of an entity. The greater tendency to omit inherent visual attributes, and the tendency to omit TYPE in conjunction with LOCATION suggest that the conceptual representation of entities is qualitatively different in these cases, giving different weight to *external* versus *inherent* properties.

3.9 Plurality and similarity

The final part of the corpus analysis considers the difference between singular and plural references, and the effect of Similarity on the production of plurals. To recapitulate the difference between levels of the Cardinality/Similarity variable, each participant in the study completed 7 references to singletons (the Singular condition), 7 to a set of two referents which had the same values on the discriminating attributes (the Plural Similar condition), and a further 6 to a set of two referents which were dissimilar on the discriminating attributes. With the exception of the manipulation of Similarity on the dimensions along which objects were discriminated, the objects were visually identical, *except that they always had different values of* TYPE. Examples of the attribute representation of referents in these conditions are reproduced below from (3.1).

(3.17) (a) (Similar)

1. $\left\{\langle \textsc{type} : \textit{fan}\rangle, \langle \textsc{orientation} : \textit{right}\rangle, \langle \textsc{colour} : \textit{blue}\rangle, \langle \textsc{size} : \textit{small}\rangle\right\}$
2. $\left\{\langle \textsc{type} : \textit{sofa}\rangle, \langle \textsc{orientation} : \textit{right}\rangle, \langle \textsc{colour} : \textit{blue}\rangle, \langle \textsc{size} : \textit{small}\rangle\right\}$

(b) (Dissimilar)

1. $\Big\{\langle \text{TYPE} : \textit{fan}\rangle, \langle \text{ORIENTATION} : \textit{front}\rangle, \langle \text{COLOUR} : \textit{blue}\rangle, \langle \text{SIZE} : \textit{small}\rangle\Big\}$
2. $\Big\{\langle \text{TYPE} : \textit{sofa}\rangle, \langle \text{ORIENTATION} : \textit{front}\rangle, \langle \text{COLOUR} : \textit{green}\rangle, \langle \text{SIZE} : \textit{small}\rangle\Big\}$

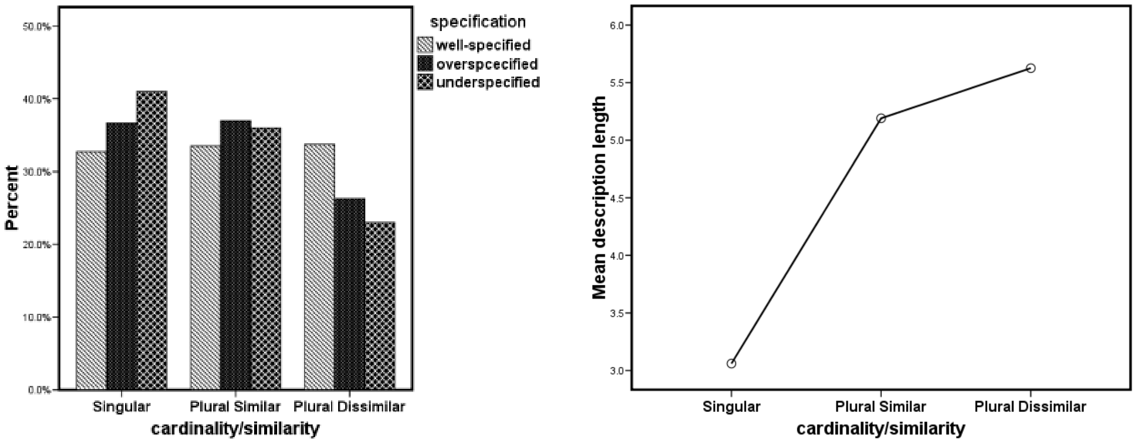
In the example, the two Similar referents have an identical value on COLOUR, whereas it is different in the Dissimilar condition. Addressing the difference among these conditions can throw some light on the extent to which authors are concise or brief when they refer to singletons and pluralities, and whether this changes as a function of the similarity of referents. It can also give some indications as to the strategy that speakers/writers follow when they refer to sets.

The Similar-Dissimilar difference has both a logical and a perceptual consequence. From a logical point of view, the minimally distinguishing description for the set in the Plural Similar condition is a conjunction. In contrast, the Plural Dissimilar condition requires a disjunctive description. Note, however, that a description in the Plural Similar condition which included TYPE would be disjunctive, unless the two objects were referred to via a superordinate term (such as *object* or *furniture item*).

Viewed from the perceptual processing angle, Similarity potentially facilitates the process of comparing the set of referents to the distractors, since the referents are identical except for TYPE, and need not be compared individually. This may result in less effort during search, or in less exhaustive search for distinguishing properties. It may also facilitate the formation of a gestalt in Pechmann's (1989) sense, since the objects are identical on all of their visual properties. As far as overspecification is concerned, this would predict (as does H3 in §3.3) that there will be no difference between Singular and Plural Similar conditions on overspecification, because there will be no greater difficulty in identifying the Similar set versus the singleton. Conversely, the Dissimilar condition potentially requires a comparison of each referent to its distractors. However, there is also the possibility that the Dissimilar condition makes the task *easier*, because referents are identical except for the contrastive attributes. Here, then, there is the possibility that H3 will be falsified: if speakers focus exclusively on the contrastive properties of the visually salient subset of the domain (the referents), they will tend to produce well-specified descriptions in this condition. If the opposite holds, and this condition incurs more search effort, requiring comparison of each referent to its distractors, then Wundt's Principle of Incrementality would predict more overspecification in this condition compared to Singulars and Plural Similar trials.

As with the analysis of different conditions in §3.7, I will use the length of descriptions as an indicator of their informativeness. The rest of the analysis proceeds as before, by looking at proportions of well-specified, overspecified, and underspecified descriptions according to the definitions in §3.6.1, and also at the likelihood of overspecification and omission of MD attributes, at different levels of the Cardinality/Similarity variable.

Figure 3.7(a) displays proportions of overspecified, underspecified and well-specified descriptions in the three Cardinality/Similarity conditions. The results have a remarkable pattern, with proportions in the Singular and Plural Similar conditions being scarcely distinguishable, except for proportions of underspecified descriptions, where Singulars have a higher rate. Both of these description types differ from the Plural Dissimilar condition, where both overspecification and underspecification was rarer. There was a significant effect of Cardinality/Similarity on proportions of well-specified descriptions, though it failed to reach significance by items ($\chi^2_1 = 12.63$,



(a) Specification of descriptions

(b) Description length

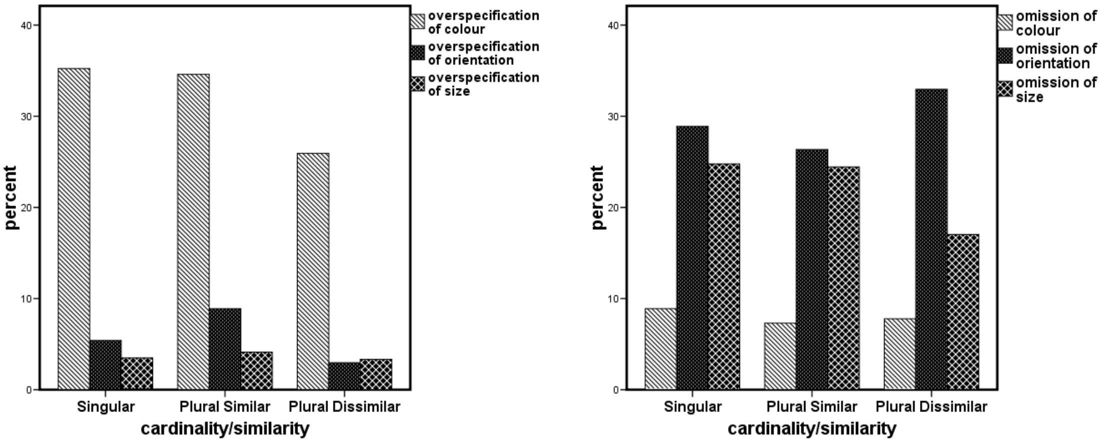
Figure 3.7: Specification of descriptions and length as a function of Cardinality/Similarity

$p = .002$; $\chi^2_2 = 2.096$, ns). The three plural types differed slightly in the likelihood of underspecified descriptions, though this only approached significance by subjects ($\chi^2_1 = 5.126$, $p = .08$; $\chi^2_2 = 1.452$, ns). Rates of overspecification did not differ between these conditions ($\chi^2_1 = 2.830$, ns; $\chi^2_2 = .125$, ns).

Pairwise comparisons were only carried out by subjects, since none of the by-items analyses reached significance. They showed that, as far as proportions of well-specified descriptions are concerned, the Singular and Plural Similar cases did not differ ($Z_1 = .777$, ns). However, the Plural Dissimilar condition differed significantly from Singular ($Z_1 = 3.157$, $p = .002$) and from Plural Similar ($Z_1 = 3.317$, $p = .001$). Similarly, Singular and Plural Similar trials did not differ on the likelihood with which people produced underspecified descriptions ($Z_1 = 1.147$, ns), while the Plural Dissimilar condition resulted in significantly less underspecification compared to Singular ($Z_1 = 2.379$, $p = .02$) and Plural Similar ($Z_1 = 2.069$, $p = .04$).

The difference in rates of underspecification and well-specified descriptions in the different Cardinality/Similarity conditions is confirmed by the analysis of description length. As Figure 3.7(b) shows, the two Plural conditions did not differ dramatically in the mean length of descriptions, but both differed from the Singular condition. The reason was that since authors preferred to use basic-level TYPE values overall, as shown earlier, they tended to produce disjunctive descriptions in the Similar case. Descriptions in the Plural Dissimilar condition were slightly longer, with a mean length of 5.6 compared to 5 for Plural Similar descriptions. This difference reached significance by subjects only ($t_1(44) = 3.119$, $p < .003$; $t_2(11) = .964$, ns), and is mostly due to the greater proportion of underspecified descriptions in the Similar condition.

Finally, I consider proportions of descriptions where MD attributes were omitted, versus those where they were redundantly included. These proportions are shown in Figure 3.8. There was a significant main effect of the Cardinality/Similarity variable on likelihood of omission of MD attributes, once again only by subjects ($\chi^2_1 = 25.117$, $p < .001$; $\chi^2_2 = 2.696$, ns). The effect of Cardinality/Similarity on overspecified usage of attributes approached significance by subjects



(a) Overspecified usage of attributes

(b) Omission of attributes

Figure 3.8: Overspecified usage and omission of attributes as a function of Cardinality/Similarity

($\chi_1^2 = 5.162$, $p = .08$; $\chi_2^2 = 1.909$, ns). Pairwise comparisons (again performed by subjects only) revealed the same pattern as with the earlier analysis of well-specified, overspecified and underspecified descriptions. The Singular and Plural Similar conditions did not differ, either in likelihood of inclusion of extra visual attributes ($Z_1 = .937, ns$) or their omission ($Z_1 = .753, ns$). In contrast, the Plural Dissimilar condition showed significantly less likelihood of unnecessary inclusion of these attributes compared to the Singular condition ($Z_1 = 4.318$, $p < .001$). The same was true of omission ($Z_1 = 2.203$, $p = .03$). The difference between the Dissimilar and Similar Plural conditions on overspecified usage was highly significant ($Z_1 = 4.089$, $p < .001$), while it approached significance on likelihood of omission ($Z_1 = 1.773$, $p = .08$). Thus, the results again confirm that the Plural Dissimilar condition is likely to result in less overspecification and less underspecification.

The results on plurals support H3 only partially. References to singletons and to sets of entities which are identical on all dimensions are no different from each other in terms of whether they are overspecified or underspecified. On the other hand, they show that the Dissimilar condition resulted in markedly less underspecification, and somewhat less overspecification as well. This may be because having two referents in focus which differed only on the contrastive attributes facilitated comparison, possibly obviating the need for comparison to the distractors. It is therefore Similarity, not Cardinality, which exerts the main influence on people's descriptive strategies.

These findings clarify some aspects of the perception and conceptualisation of multiple referents, especially the extent to which visual (dis)similarity of referents which are in the focus of attention (because they are visually salient) allows a contrast to be made between them, facilitating the formulation of a linguistic message in which precisely those properties which are contrastive are included. This will play a role in the GRE evaluation that is to follow.

3.10 Summary and outlook

This chapter began by introducing the notion of semantic transparency as a desideratum that corpora for the study of reference should satisfy. The TUNA Corpus, which was described in the first half of the chapter, is an example of such a corpus. Apart from semantic transparency, this corpus is distinguished from many other linguistic resources in that it is the result of an experimental study which also sought to balance the materials used and cover a sizable portion of the space of possibilities allowed by the hypotheses that it was designed to address. The annotation of the corpus, which was made possible by the prior availability of the domains in which descriptions were elicited, also serves as an example of how semantic transparency can be achieved to make a resource machine-readable.

This chapter also reported on the results of an empirical study on the corpus that shed light on three principal issues:

1. The impact of the perceived fault-criticalness of a referential communicative task. Here, the results suggest that though fault-criticalness may play a role in the extent to which people over- or underspecify their descriptions, this interacts in a significant way with whether or not they use locative expressions to describe entities. I have proposed that the reason for the apparent reduction in the use of inherent visual attributes of objects when LOCATION is also present in a description is due to the interaction of two sub-systems in the perceptual processing of objects. First, orienting towards an object implies processing its location; second, once this is carried out, analysing the object's visual features and constructing a Gestalt representation can proceed. In case an object is identifiable in terms of its location, the second of these processes is more likely to be interrupted. Though based on some insights from studies and models of visual attention, this explanation remains speculative, since only through online processing studies can it be confirmed.

2. The preferences evinced for certain attributes over others in people's descriptions. This is the cornerstone of psycholinguists' explanation of the tendency to overspecify and is the motivating observation for the Gestalts Hypothesis of Pechmann (1989), which has served as the underlying hypothesis for the present work. The results of the present study, based not only on overspecification but also underspecification data, show that this hypothesis is largely correct and extends previous results to other attributes, such as ORIENTATION. An interesting outcome of the study was the way that an increased likelihood of overspecified usage of an attribute was usually mirrored by a corresponding decrease in likelihood of omission when it was not strictly required for identification. This picture is complicated somewhat by the apparent discrepancies in the data for ORIENTATION, which was more likely to be overspecified than, say, SIZE but also more likely to be omitted. I have proposed that this is due to a difference in the difficulty of processing certain values of ORIENTATION. Another complicating factor (one that is reminiscent of earlier results by Arts (2004)) is the apparent mutual dependency of different values of an attribute. This is the case with LOCATION, where the vertical dimension is more highly preferred, while the use of the horizontal dimension may be dependent on the use of the vertical.

3. The difference between singular references and plural references to two objects, together

with the effect of similarity in reference to sets of objects. Two important results emerge from this part of the study. First, the trends related to overspecification and attribute preferences carry over from the singular to the plural case. Second, perceptually similar objects are described by humans using strategies that are very similar to the way they describe singletons, whereas two target referents which are perceptually dissimilar are less likely to result in an overspecified or underspecified description.

With these results in the background, I now move on to an evaluation of the GRE algorithms introduced in Chapter 2. The evaluation will serve two main functions. It constitutes one of the first systematic evaluations of these algorithms against human data, and certainly the first to consider plurals as well as singular descriptions. Moreover, the performance of the algorithms compared to the corpus data serves as a further test of the hypotheses tested in this chapter. Because the evaluation explicitly tests algorithms on plural descriptions, it will highlight their limitations on such domains, and provide some motivation for the work reported in the second part of this thesis (Chapters 5–7).

Chapter 4

Evaluating GRE algorithms against a semantically transparent corpus

4.1 Introduction

This chapter uses the corpus described in Chapter 3 to evaluate the Incremental (IA), Greedy (GR) and Full Brevity (FB) content determination heuristics introduced in Chapter 2. Because of the nature of the corpus and the way it was annotated, it is possible to expose the algorithms to the same domains as the authors in the corpus, and compare their descriptions of the same targets to the human-authored descriptions. Moreover, the empirical study of the previous chapter will serve to motivate some of the practical decisions that need to be made when setting the external parameters of algorithms. This is especially relevant for the Incremental Algorithm which, as I argued in (§2.7, p. 47) characterises a *family* of algorithms in a given problem space, each defined by a different preference order.

The rest of this chapter is organised as follows. I begin by comparing the rationale of the present evaluation to that of other forms of evaluation in NLG. In §4.2 (p. 103), I give an overview of previous evaluation studies in GRE. This serves to highlight a number of unanswered questions in previous studies which the present chapter seeks to address. §4.3 (p. 107) describes the implementation of the algorithms used in the evaluation, conforming to the formal definition of a GRE problem instance given in Chapter 2. After describing the evaluation metric used (§4.3.2, p. 110), I go on to discuss the versions of the IA that were implemented for the evaluation. This is a particularly important issue, because the number of possible IAs grows exponentially in the number of available attributes in a domain. Moreover, comparing a huge number of algorithms impacts the statistical reliability of the results. Therefore, the results of the empirical study in the previous chapter were used to inform the decision on which preference orders to test.

The results of the evaluation are reported in §4.4 (p. 113)–§4.6 (p. 123). This evaluation sought to investigate three central issues:

1. **The relative performance of the Incremental Algorithm** against its predecessors, namely the Greedy and Full Brevity heuristics (§4.4, p. 113). In all the results reported here, there is a version of the IA which outperforms the two earlier interpretations of Gricean Brevity. This is exactly as predicted by the previous empirical study and by psycholinguistic evidence. However, this observation is inseparable from the answer to the next question.

2. **The impact of different preference orders on the** IA (§4.5, p. 120): It turns out that in

every test conducted to compare the IA to GR and FB, there is also at least one order whose performance is either worse, or not better than that of FB and GR.

3. **The performance of all these algorithms on singular versus plural data** (§4.6, p. 123): Here, the results show clearly that when these algorithms are extended beyond their original remit to deal with disjunction and plurality, their performance declines dramatically.

4.1.1 Rationale

The method adopted here is based on the assumption that in evaluating the output of NLG modules, human output in comparable situations can be used as a 'gold standard'. However, in the present context, this assumption needs to be qualified. There is no guarantee that the descriptions in the TUNA Corpus are 'perfect' or even 'sufficiently adequate' as identifying descriptions. Ascertaining this would require a hearer-oriented study using the human-authored descriptions, testing the extent to which they facilitate a hearer's identification task. By contrast, the present study aims to compare the content determination decisions made by humans to those made by algorithms. Although 'imperfections' exist in the data (e.g. a small proportion of descriptions in the corpus were found to be underspecified), the results of the data analysis also showed that there are clear, non-random trends in the way people select content, and that these trends conform to earlier results and extend them.

Other forms of evaluation, sharing this rationale to a greater or lesser extent, are conceivable in NLG. For instance, in a **task-oriented evaluation**, it is the extent to which the output of an NLG system achieves its communicative goal that serves as a test of the viability of the technology, and the effectiveness of the algorithms it incorporates. A good example of this is the large-scale clinical trial used to test the effectiveness of a system designed to generate personalised smoking cessation letters (the STOP system; Reiter et al., 2003). In this test, the measure of success of the system was whether or not a significant proportion of people who received such automatically generated personalised letters would stop smoking, compared to populations who had either not received personalised letters, or whose letters had been manually generated. In GRE, a task-oriented evaluation might measure the extent to which automatically generated descriptions enable listeners or readers to successfully identify an object, thereby taking a reader's perspective. In contrast, in the present evaluation I adopt a speaker's perspective. The possible differences between the two are worth emphasising, in view of the partial evidence for speaker-listener asymmetries discussed in §2.6.3 (p.44).

There is a fairly well-established evaluation tradition in the Natural Language Processing literature that uses corpus-based evaluation metrics. Such metrics have come to dominate the literature on Machine Translation (e.g. Papineni et al., 2002); recent work in NLG has also evaluated output against domain-specific human-produced corpus texts (e.g. Barzilay and Lapata, 2005; Lapata and Barzilay, 2005; Belz and Reiter, 2006). These studies, which focused on issues pertaining to both content determination and realisation, used parallel corpora, wherein texts are coupled with the corresponding dataset or domain (the 'semantics'). Even when such corpora are available, the extent to which they should be treated as gold standards has been questioned (Reiter and Sripada, 2002b). Because such texts are often produced by a variety of authors, there is considerable variation in the mapping from content to NL, partly because of individual variation (Reiter and Sripada, 2002a), and partly because the circumstances in which such texts are produced are not always controllable. Thus, Belz and Reiter (2006) found that expert weather forecast readers tended to rank corpus forecasts, which were produced by fellow experts, lower than those generated automatically. Moreover, there was a significant mismatch between automatic evaluation scores for generated texts based on a comparison with corpus texts, and the scores assigned by human readers to the same texts. The authors conclude that 'if an imperfect corpus is used as the gold standard' for automatic corpus-based evaluation metrics, 'then high correlation with human judgements is less likely' (Belz and Reiter, 2006, p.318). A related point was made by van Deemter (2004), who raises the possibility of an algorithm matching only a small subset of the individuals whose output makes up the totality of observations in an experimental dataset or corpus. This would presumably result in a low overall algorithm-human match, but arguably does not warrant the conclusion that the algorithm performs poorly. A different, though related, question when examining algorithms against a corpus is the source of variance for the analysis. An algorithm may turn out to be better than others when its performance is averaged over a set of domains (a 'by-items' analysis), but have a low average match to a set of authors (a 'by-subjects' analysis). Taking both perspectives is only possible in a corpus which, in addition to satisfying the transparency requirement, is also balanced, in the sense that each 'item' (domain) is represented an equal number of times, and authors were exposed to comparable sets of domains varying along well-defined parameters.

In the case of GRE, using corpora is tricky for similar reasons. First, because of the semantically intensive nature of the GRE task, the output of GRE algorithms depends to a large extent on the way the semantics of definite descriptions are handled, given a well-defined domain. Corpora texts seldom contain a domain representation, and NL utterances in corpora are morpho-syntactic realisations from which the semantics has to be inferred. Problems specific to reference also arise with text corpora. Consider once more the example of the painting *Las Meninas* from Chapter 1. Suppose a number of texts produced by different authors were collected and mined for definite descriptions referring to different parts of the painting. Despite the visual representation of the painting, assumptions about domain representation cannot be straightforwardly made. Different authors may have different levels of knowledge about the painting, enabling the use of content that is not accessible to other authors. Conversely, there may be aspects of the painting that the authors did not use in their descriptions; these correspond to properties or attributes that were not included in their process of 'content determination', but in order for an appraisal of GRE algorithms to be empirically well-founded, such attributes would have to be included in the domain representation. The availability of a semantically transparent corpus, with an explicit domain representation paired with human descriptions annotated at a semantic level, goes some way towards resolving these problems, by making the input to a GRE algorithm as similar as possible to that of the human author.

Before turning to details of the evaluation, it is worth giving an overview of the few previous studies – three in all – that have compared and evaluated GRE models against human data. This will also permit the formulation of the research questions to be addressed here.

4.2 Previous GRE evaluations

Considering the influence of the work of Dale and Reiter (1995), the lack of empirical evaluation in GRE is surprising. The three recent exceptions to this tendency are corpus-based, though they differed in their principal goals. Jordan and Walker (2000, 2005), and Gupta and Stent (2005) compared the output of the IA to other models in a dialogue context. In the former study, the aim was to assess the performance of a referential strategy designed principally for identifying referents, to other models which take other pragmatic factors into account. The study by Gupta and Stent (2005) also introduced other pragmatic factors when evaluating the IA. The third study, by Viethen and Dale (2006) compared the IA and GR algorithms, as well as the algorithm for the generation of relational expressions by Dale and Haddock (1991), in a non-dialogue setting using a small corpus that was constructed for the purpose.

4.2.1 Jordan and Walker (2005)

Jordan and Walker (2005) was a larger-scale replication of a previous study (Jordan and Walker, 2000), focusing exclusively on the dialogues in the COCONUT corpus, a collection of task-oriented dialogues in which interlocutors had to resolve a joint task (buying furniture on a fixed budget) in a well-defined domain. The authors used a machine learning paradigm to compare attribute selections for definite descriptions made by the IA against two other models:

1. The *Intentional Influences* (II) model focuses mainly on the multiple intentions that human referential descriptions seek to satisfy. This is an empirically-informed model, based on studies of the COCONUT corpus itself (Jordan, 2000b, 2002). Although not originally proposed as a formalised algorithm, II has been implemented and evaluated on at least one other occasion (Jordan, 2000a), although the precise details of the implementation are likely to vary from domain to domain (in particular, one needs a predefined set of 'intentions' to be taken into account by the algorithm).

2. The *Conceptual Pacts* (CC) model of Brennan and Clark (1996) holds that dialogue partners 'agree', or converge on, the set of lexical items used to refer to an entity, meaning that a person's content determination is influenced by their interlocutors' references. This model is *not* an algorithm, but an empirically-informed theoretical framework for interpreting human dialogue interaction, falling under the rubric of Clark's (1996) *language-as-action* paradigm.

The method used by the authors was as follows:

1. Annotation of a sample of COCONUT dialogues with features pertaining to the three models. For this purpose, descriptions were annotated as a set of classes with input features that were used as predictors for those classes.

2. Running the RIPPER machine-learning algorithm (Cohen, 1996) to learn content selection rules from the annotated data. RIPPER's output consists of production-like IF-THEN classification rules based on abstractions over input patterns, with a statistical weighting that determines their precision and recall on a corpus of unseen test data;

3. Comparing the coverage of the resulting rule sets obtained from the three models.

To test the different algorithms, different sets of features were combined. Thus, the CC model primarily requires information about previous lexical choices in references to an object, while II depends on information about the partners' state of agreement, their commitment to proposals and so forth. For the IA, the principal source of information was the available attributes, the distractors, and salience information to calculate the context set. Somewhat surprisingly, Jordan and Walker (2005) take this to be the central aspect of the IA, whereas the actual determination of the context set was not delved into in any great detail by Dale and Reiter (1995), who assumed that this was the set of objects currently being attended to.

The outcomes of the study revealed that all three models performed significantly above a baseline, with an improvement in performance of the II (42.4%) over the IA (30.4%) and CC (28.9%). However, the best coverage (ca. 60%) was achieved by combining the II and the IA. This result seems to suggest a strong dependency on the domain of discourse. The relatively good performance of the IA means that identification (which is the principal communicative intention modelled by the algorithm) plays a crucial role even in a task-oriented dialogue setting. However, the complex interplay of factors in COCONUT, where there are very clearly-defined constraints on what interlocutors must achieve, means that identification is only one of several communicative intentions. Thus, it is hardly surprising that II outperformed IA, since the former explicitly models the intentions that the corpus was designed to address. Perhaps it is their joint overall improvement that constitutes the most interesting outcome of the study.

This study therefore leaves open the question as to *whether the* IA *is an adequate model of referent* identification *compared to alternative models*, a question that was not central to this study, which compared the IA to models that included further communicative intentions. Jordan and Walker's evaluation makes a strong case for taking intentions beyond identification into account when building computational models (see also the discussion in §2.6.3, p. 44). There are, however, two additional problems that call for caution when generalising the results. First, they are intimately bound to the nature of the data. Because of the high dependency on annotations at various levels in the corpus, it is possible that the performance of the different models would not generalise to a new dataset, involving a different communicative task or a different domain of discourse[1]. Second, as observed earlier in this section, the models with which the IA was compared have never been formalised to the same degree as the IA. Thus, there is a sense in which the annotation that formed the basis for the evaluation was an *interpretation* of these models[2].

4.2.2 Gupta and Stent (2005)

The study by Gupta and Stent (2005) – again on task-oriented dialogues – was carried out using data from COCONUT and the MAPTASK corpus (Anderson et al., 1991). In MAPTASK the interlocutors' task was to converge on a common route through a map containing several named landmarks. The evaluation compared the IA to a version of the Greedy algorithm by Siddharthan and Copestake (2004, SC), against a baseline procedure that included the TYPE of an object, and randomly added further properties until a referent was distinguished. Additionally, each algorithm was combined with two dialogue-specific models: (a) a version which reordered the preference order of

[1] This problem is acknowledged by the authors, and is raised here in the spirit of a cautionary note about the present evaluation.

[2] Note that this is not a critique of the level of agreement reached among annotators.

attributes based on the last description in the dialogue; (b) a version which re-used properties in the last description of the target. These additions were meant to address Brennan and Clark's (1996) *Conceptual Pacts* (CC) model. Unlike Jordan and Walker, this model was incorporated as an add-on to existing algorithms, rather than evaluated separately against them. In addition to these versions, Gupta and Stent coupled the algorithms with different procedures for the realisation of modifiers in the generated NPs; thus, the evaluation took both content determination and some aspects of syntactic realisation into account, using a single evaluation metric which combined:

1. The attributes included by the algorithm and a human author;
2. The attributes included by an author but omitted by the algorithm;
3. The attributes included by the algorithm but omitted by an author;
4. Whether the attribute was placed in a syntactically correct position by the algorithm, compared to an author's description.

Three major points emerge from this study, two of which are related to those raised by the Jordan/Walker evaluation. The first is again the strong dependency of the outcomes on the nature of the corpus. The baseline algorithm outperformed both IA and SC algorithm on the MAPTASK data. This is because most of the referents in the task were landmarks on a map, and their TYPE attribute had the status of a proper name (there were no 'real' distractors). By contrast, the COCONUT domain is more elaborate, and the two algorithms outperformed the baseline on this corpus. The second point has to do with the nature of the discourse in which algorithms are being evaluated. On the COCONUT dialogue data, the original IA was outperformed by both variants that took into account partner-specific effects, exactly as previous empirical work on this corpus would predict. There is a potential confounding factor at work here, since identification is often not the only referential goal of interlocutors.

Finally, the evaluation of the IA and SC – both designed for a 'purely semantic' content-determination task – by also taking into account syntactic factors such as modifier placement, is questionable. Neither of these procedures incorporates a syntactic realisation module, but since this was included in the scores used to evaluate the algorithms, the contribution of the original algorithms to the overall match with human data is somewhat obscured.

So far, empirical studies that compare the IA to alternative strategies have either not tested the IA on an equal footing with those models, or, in the case of Gupta and Stent (2005), have obscured the contribution of the content determination algorithm proper by taking other factors into account. The most serious shortcoming of the two studies reviewed so far, however, is that neither makes explicit the way the preference order of the IA was determined for the experiments. As argued in §2.7, different preference orders potentially result in completely different algorithms. Clearly, feasibility issues may deter the experimenter from carrying out an exhaustive comparison of $n!$ versions of the IA in a domain with n attributes; however, in order to properly test the model, an explicit account of how the preference order was determined is crucial.

4.2.3 Viethen and Dale (2006)

In comparison to the studies discussed above, the one by Viethen and Dale (2006) was a much more straightforward comparison of the IA and other algorithms. Here, I focus exclusively on one aspect of the evaluation, the comparison between the IA and GR[3].

In their methodology, Viethen and Dale stuck to the identification criterion as the sole communicative intention, setting up a semi-formal experiment in which participants were asked to refer to drawers (of their own choice) in a filing cabinet. The drawers in the cabinet differed on the basis of COLOUR and LOCATION, the latter consisting of the three separate attributes of ROW, COLUMN and, in case the drawer was in a corner, CORNER-HOOD. Since TYPE was the same for all referents, it was not taken into account in the evaluation. The experiment resulted in a small corpus of descriptions ($N = 118$, of which 103 were non-relational). The comparison of IA and GR revealed that GR had a recall rate of 79.6%, compared to a 95.1% rate for the IA, both figures excluding relational descriptions[4]. Moreover, the corpus contained a relatively small number of overspecified descriptions (29 out of 103). Of these, the IA reproduced all but five.

No figures are provided to indicate the degree of variance between individuals in this study, and recall was calculated over the corpus as a whole, without taking into account the degree of match between the algorithm and particular subjects. Although the results seem favourable for the IA, they represent an average over all $(4! =)$ 24 possible preference orderings for the IA. In addition, GR was combined with a preference order to resolve ties between alternatives with equal discriminatory power.

Making an overall comparison of 24 different versions of an algorithm obscures the the status of the recall percentage, since this amounts to a score obtained by 24 different algorithms (or 24 different statistical hypotheses) in tandem, and the extent to which a given version of the IA, with a particular preference order, contributes to the overall rate is unknown. Although the methodology does away with a lot of the additional factors of the previous two studies, it still does not permit an answer to the question of whether the IA, compared to at least one other alternative algorithm, is a better model.

4.2.4 Interim summary

The three studies reviewed here were among the first to systematically compare the gold standard Incremental Algorithm to alternative models. Because the studies were frequently designed with more than one evaluation goal in mind, and because they were either conducted on data that required going beyond the IA itself, or they obscured the way the IA's preference order was determined, the results are not a reliable estimate of the algorithm's performance. They raise a number of methodological issues in GRE evaluation:

1. The generalisability of evaluation results, as a function of the corpus used in a study, and the compatibility between that corpus and what the algorithms tested were actually designed to do.

2. The method of comparison, in particular, whether only semantic factors are taken into account, or whether realisation issues should also play a role.

3. The method of evaluation of an algorithm whose parameters may radically change its behaviour (this is especially true of the preference order in the IA).

[3] The authors actually claim they are comparing the IA and FB, but it is the greedy version of Gricean brevity that they formalise. In addition, they compare the two algorithms to Dale and Haddock's (1991) relational algorithm.

[4] *Recall* was defined by the authors as the number of descriptions in the corpus which an algorithm reproduced perfectly.

4. The question of how to actually compare algorithms and human data. Taking the data as a corpus of observations – a *by-items* analysis – gives a good indication of the overall match to a set of instances, but does not take into account variance as a function of individual variation. It is conceivable that an algorithm be a good match to a few individuals, but not others.

5. The question of how far to control communicative intentions over and above the identification goal that IA and related models were designed to address.

Finally, some crucial evaluation issues that address the claims that have been central to GRE research over the past decade remain unaddressed:

Q1 To what extent does the IA approximate human output relative to FB, and GR?

Q2 To what extent does the performance of the IA depend on preference orders?

Q3 Does the extent to which algorithms like the IA compare to humans differ as a function of the type of descriptions? In particular, do extensions of the algorithms to plural descriptions involving disjunction, and possibly other expression types, such as gradables, change the overall picture?

The study reported in the next few sections sought a reply to these questions.

4.3 Evaluating the algorithms

For the evaluation study, I used the corpus to automatically compare the output of IA, GR and FB to the human-produced descriptions. These were represented as logical forms compiled from the corpus annotations, based on the `DESCRIPTION` and `ATTRIBUTE` tags, using the rules described in §3.5 (p. 78). The evaluation focused on the degree of agreement between algorithms and human authors on attribute selection.

Since some corpus domains allowed authors to use LOCATION and some didn't, the corpus data was divided into two sets. The $+$LOC dataset contained the descriptions in the $+$FC$+$LOC condition, as well as those descriptions in the other conditions on which participants had used LOCATION. Due to some system errors, location information was missing from some of the domain representations[5]. These were omitted from the evaluation. The final dataset consisted of 412 descriptions from 26 authors. The other dataset contained all the other descriptions, from the $-$LOC conditions, for a total of 444 descriptions from 27 authors[6]. This is the simpler of the two datasets, with knowledge bases comprising only 3 attributes, apart from TYPE. The $+$LOC dataset is more complex, and offers more scope for variation. It has 5 attributes that can be used to distinguish the referents, because X-DIMENSION and Y-DIMENSION are included. Moreover, since the position of objects was determined separately for each domain and each participant, the behaviour of the algorithms is not immediately predictable from the KB, as it is in the $-$LOC dataset. For instance, Y-DIMENSION may well be minimally distinguishing for the set of referents (hereafter denoted R)

[5] This did not affect the empirical study of Chapter 3, since the descriptions still contained the annotations required for analysis.

[6] The number of individual authors in the two datasets does not sum to 45 because the descriptions by authors who had used LOCATION in the $-$LOC conditions were added to the $+$LOC dataset

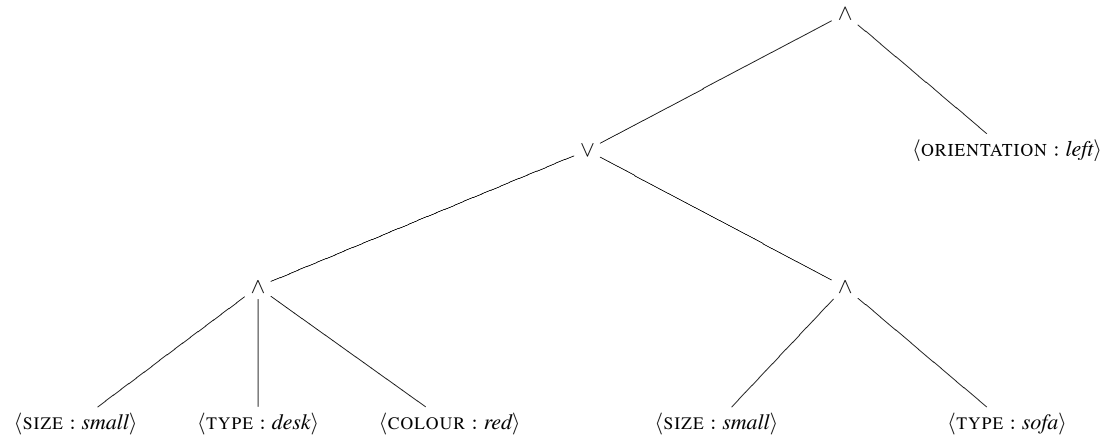


Figure 4.1: Tree representation of the formula in (4.1)

in one domain offered to one author, but not in that of a different author, even when the domains represent the same experimental conditions.

In the −LOC dataset, only GR and IA were compared, because FB gives identical results to GR. Although the discussion in §2.5 (p. 33) showed that FB and GR cannot be guaranteed to yield identical outputs unless only one property is required to identify a referent, their identity in the −LOC cases emerges as an artifact of the experimental design in the TUNA Corpus. Recall that in any trial, the minimal description MD was calculated such that there was no literal that was true of the intended referent which had greater discriminatory power than the literals making up the minimal description. Therefore, GR would always select the properties making up MD, while FB would return MD by definition. This calculation only took the inherent visual attributes of objects into account, so that the identity of FB and GR no longer holds in the +LOC data, which also includes Y-DIMENSION and X-DIMENSION.

4.3.1 Implementation of the algorithms

The implementation of the algorithms used in this study forms part of the GRE-API, a Java package for the Generation of Referring Expressions, constructed as part of the TUNA Project. This section focuses on some relevant aspects of the knowledge representation component of the API, which has a bearing on how the similarity or degree of agreement of two descriptions is computed.

As per the standard view, each generation algorithm is coupled with a KB, with KB properties represented as attribute-value pairs. Formulae – that is, disjunctions and conjunctions of properties – are represented as unordered trees, where non-terminal nodes are logical operators and attribute-value pairs are the leaf nodes. An example of a tree representation of the formula in (4.1) is shown in Figure 4.1.

(4.1) $$\begin{gathered}\Big[\,(\langle\text{SIZE} : \textit{small}\rangle \wedge \langle\text{COLOUR} : \textit{red}\rangle \wedge \langle\text{TYPE} : \textit{desk}\rangle) \vee (\langle\text{SIZE} : \textit{small}\rangle \wedge \langle\text{TYPE} : \textit{sofa}\rangle)\,\Big] \\ \wedge \\ \langle\text{ORIENTATION} : \textit{left}\rangle\end{gathered}$$

The Figure shows the format of both the corpus descriptions, after compilation from the XML annotation, and the algorithm output. After compiling a corpus description, the evaluation program read in the domain data from the corpus XML file, populating the KB with the relevant properties, and then ran the algorithms on the same domain. This meant that humans and algorithms were exposed to the same Knowledge Base. To avoid penalising an algorithm for the omission of an attribute which had been included in a corpus description, but which had the value `other` (e.g. an `other` value of TYPE in the description *the picture in the top left*), only attributes (as opposed to values) were considered by the evaluation function used.

The implementation of the three algorithms in the GRE-API essentially conforms to the formalisation given in Chapter 2, placing all algorithms on a common footing. To recapitulate, each algorithm is implemented with the following components:

1. A priority queue Q which imposes an ordering on properties in the search space depending on the ordering relation underlying the algorithm. This is implemented as a dynamic queue, in which ordering can change as the state of the algorithm changes (see §2.4.1, p. 32);

2. Two queueing functions associated with the priority queue: $dequeue(Q)$ returns the highest property in the priority queue at a given state, and $enqueue(p, Q)$ creates combinations of properties (disjunctions and/or conjunctions) involving p, adding them to the queue;

3. A main function which takes as input the set R of intended referents, the contrast or distractor set C and the set of relevant properties P_R (see §2.4.1 for a definition of these terms). The function begins by enqueing all the relevant properties P_R, and proceeds by dequeuing a property, checking for its contrastive value against the distractor set C, and updating the description accordingly. The function terminates on success, or when the queue has been exhausted.

As discussed in Chapter 2, the main difference between the algorithms is the way ordering of properties in Q takes place. For instance, the priority queue associated with FB gives priority to shorter combinations. This algorithm searches through all combinations of properties until a distinguishing description is found. For the other two algorithms, the main function dequeues a property, tests whether it is true of R and has some discriminatory value, updates the description accordingly, and, if the domain is plural, enqueues disjunctions involving the property if the description is not distinguishing, and the queue has not been exhausted (cf. Bohnet and Dale, 2005, for a related view). Besides GR, FB and IA, a baseline algorithm was included in the evaluation, which randomly added properties which were true of an intended referent to a description until it was distinguishing (cf. Gupta and Stent, 2005). This is hereafter referred to as RAND.

Since the corpus contains a large number of plural descriptions, each of the algorithms was extended to deal with disjunction[7] in the way proposed by van Deemter (2002). Thus, suppose that a property p were currently under consideration, and the property were included in the description. If the description were not distinguishing, the algorithm would enqueue p disjoined from every other property in $P_R - \{p\}$. Disjunctions pose a potential problem for the IA because they cannot be straightforwardly distinguished on the basis of the attribute order. For this reason, my implementation generalised the notion of ordering as follows. Let F be an arbitrary formula,

[7] Negation was not at issue, since the corpus only contained 2 instances of a negated property. Therefore, negation was never used by the algorithms in this study.

$|F|$ be the number of properties in F, and $att(F)$ the set of attributes in F. The relative ordering of two formulae, based on a preference order PO, is generalised from that given in §2.7 (p.47) as follows:

$$F >>_{p_{IA}} F' \leftrightarrow \begin{cases} \sum_{\text{A} \in att(F)} index(\text{PO}, \text{A}) < \sum_{\text{A}' \in att(F')} index(\text{PO}, \text{A}') \text{ if } |F| = |F'| \\ |F| < |F'| \text{ otherwise} \end{cases} \tag{4.2}$$

In other words, properties are ordered by length (shorter properties first), whereas, if two formulae are of the same length, one takes precedence if the sum of indices of the attributes it contains, relative to $\mathcal{PO}$, is less than the sum of indices of attributes in the other formula. This ordering relation will order literals with respect to each other in the familiar way, and will order literals before disjunctions of length 2 or more. For disjunctions of the same length, it will, for example, order ⟨COLOUR : *red*⟩ ∨ ⟨COLOUR : *green*⟩ before ⟨COLOUR : *red*⟩ ∨ ⟨ORIENTATION : *backward*⟩, if COLOUR is placed earlier than ORIENTATION in PO.

Because human authors consistently used TYPE in their descriptions, each algorithm included this attribute irrespective of its discriminatory value. This was recommended by Dale and Reiter (1995) for the IA, and was also added to FB and GR to avoid penalising their performance unnecessarily. Furthermore, because LOCATION (X- and Y-DIMENSION) is numerically represented in the corpus domains, the treatment of this attribute used the algorithm for gradable properties by van Deemter (2006). This was implemented as a preprocessing stage, in which gradable properties were transformed into inequalities. Following each run of an algorithm, the inference rules proposed by van Deemter and described in §2.7.4 (p. 55) were applied.

4.3.2 Comparing descriptions

Logical forms derived from the human-produced data were compared to those generated by the algorithms using the Dice coefficient, the same evaluation metric as was used for the evaluation of inter-annotator agreement in the corpus study of Chapter 3 (p. 82). Its calculation is reproduced in equation (4.3), where D and D' are two descriptions, and $att(D)$ the set of attributes in a description.

$$dice(D_1, D_2) = \frac{2 \times |att(D_1) \cap att(D_2)|}{|att(D_1)| + |att(D_2)|} \tag{4.3}$$

Once again, the rationale for using this measure is that the evaluation focuses primarily on the degree of match of two descriptions on their content. Because properties could appear more than once in a disjunction, descriptions were again represented as multisets rather than sets, so that the coefficient took into account each individual occurrence of an attribute. As a result, the function can reflect differences arising from one description having more 'epistemically redundant' properties than another, for example because one description contains an occurrence of the same attribute more than once (cf. Gardent, 2002). As an example, consider the disjunctive description (4.4a) . This would be compiled into the logical form (4.4b). Suppose an algorithm produced the description (4.4c).

(4.4) (a) the small red sofa and the large green desk

(b) $$(\langle \text{SIZE} : \textit{small}\rangle \wedge \langle \text{COLOUR} : \textit{red}\rangle \wedge \langle \text{TYPE} : \textit{sofa}\rangle)$$
$$\vee$$
$$(\langle \text{SIZE} : \textit{large}\rangle \wedge \langle \text{COLOUR} : \textit{green}\rangle \wedge \langle \text{TYPE} : \textit{desk}\rangle)$$

(c) $$(\langle \text{SIZE} : \textit{small}\rangle \vee \langle \text{SIZE} : \textit{large}\rangle) \wedge (\langle \text{COLOUR} : \textit{red}\rangle \vee \langle \text{COLOUR} : \textit{green}\rangle)$$
$$\wedge$$
$$(\langle \text{TYPE} : \textit{sofa}\rangle \vee \langle \text{TYPE} : \textit{desk}\rangle)$$

Despite their different logical structures, these descriptions would have a Dice coefficient of 1, indicating perfect agreement on content. Since Dice, as used here, abstracts away from the structure of logical forms and operators in descriptions, a description produced by an algorithm has a score less than 1 if either (a) it contains properties not included in the human-authored description; or (b) it omits properties included therein. Thus, formulae which are non-identical in terms of their logical structure, such as the ones exemplified above, can sometimes come out as identical in terms of content. In this manner, the focus of evaluation remains restricted to the attribute selection of the algorithms in comparison to humans. A different evaluation measure is used in the following chapter, where two strategies for plural generation are compared.

4.3.3 Determining preference orders for the IA

One of the questions I set out in §4.2.4 (Q2) was related to the impact of different preference orders on the IA. Taking all possible orderings into account can result in a combinatorial explosion. This would not only make analysis difficult, but would also make it statistically less meaningful. With the furniture domain, there are at most 5 attributes (the three MD attributes and two LOCATION attributes), not including TYPE, which yield 120 different orders. If each order were compared to every other one, this would yield $\binom{120}{2} = 7140$ pairwise comparisons. With so many statistical tests, the probability $(1-\alpha)$ that a true null hypotheses will be rejected at the α level of significance increases enormously[8]. For this reason, I followed a different strategy to examine the impact of different orders.

The data analysis in the preceding chapter showed up some preferences, notably for COLOUR and LOCATION, especially in the vertical dimension. The relative ordering between SIZE and ORIENTATION was not clear-cut, due to conflicting overspecification and underspecification data. In the $-$LOC dataset, there are two possible orders with COLOUR first. To these, I added a **baseline order**, which reverses the trends observed in the data. The three orders are shown in (4.5).

(4.5) (a) C >> O >> S
(b) C >> S >> O
(c) S >> O >> C (baseline)

If preference order really has an impact on the performance of the IA, and given the results of the corpus analysis, the first two of these should not differ from each other, but the baseline order should result in a significant decline in performance.

For the $+$LOC dataset, the following assumptions were made about preference ordering, once again based on the data analysis:

[8] In general, the likelihood of accepting a null hypothesis in conducting n tests at an α level of significance is $(1-\alpha)^n$. Thus, the likelihood decreases with the number of tests. See Bland and Altman (1995) for discussion.

1. Y-DIMENSION always precedes X-DIMENSION. This is motivated by the data reported in §3.8.1 (p. 91), in which Y-DIMENSION was more frequently used in isolation than X-DIMENSION. This, in combination with the fact that they tended to be used together most often, suggests that the use of X-DIMENSION was more likely when Y-DIMENSION was also present (cf. Arts, 2004).

2. Numeric values of X-DIMENSION and Y-DIMENSION are compiled into inequalities. This means that the extension of a property of the form $\langle \text{A} > n \rangle$ will include the union of all properties of the form $\langle \text{A} = m \rangle$ where $m < n$ in the KB. Thus, no single locative property is guaranteed to be minimally distinguishing. Since people often used both dimensions in their descriptions (cf. §3.8.1, p. 91), this increases the likelihood of both location dimensions being considered by an algorithm. Moreover, whether FB or GR select a locative attribute will depend on the discriminatory power of the logically strongest inequality. This will change from domain to domain, and therefore increase the variation in the results.

3. Within a numeric attribute A, inequalities of the form $\langle \text{A} > n \rangle$ are always ordered before properties of the form $\langle \text{A} < m \rangle$. Properties are also ordered by logical strength. Thus, $\langle \text{A} > n \rangle$ precedes $\langle \text{A} > n' \rangle$ iff $n > n'$. Similarly, $\langle \text{A} < m \rangle$ precedes $\langle \text{A} < m' \rangle$ iff $m < m'$.

With these restrictions, the number of possible preference orders decreases from 120 to 20. These are constructed by taking the two orders in (4.5), excluding the baseline order, and interpolating X- and Y-DIMENSION, observing the above restrictions. The first 10 resulting orders are shown in (4.6). The other 10 are obtained by switching O(rientation) and S(ize).

(4.6) (a) Y >> X >> C >> S >> O

(b) Y >> C >> X >> S >> O

(c) C >> Y >> X >> S >> O

(d) Y >> C >> S >> X >> O

(e) C >> Y >> S >> X >> O

(f) C >> S >> Y >> X >> O

(g) Y >> C >> S >> O >> X

(h) C >> Y >> S >> O >> X

(i) C >> S >> Y >> O >> X

(j) C >> S >> O >> Y >> X

Once more, a baseline order suggests itself from the data analysis. This time, it is COLOUR *and* Y-DIMENSION that are placed at the bottom of the list. Of the possible orders with this restriction, I selected the one in (4.7).

(4.7) X >> O >> S >> Y >> C (baseline)

Hereafter, I adopt the convention of referring to an instance of the IA by the initials of its preference order (e.g. COS). Baseline orders are referred to as IA-BASE. To evaluate the different algorithms, I compare them in pairwise fashion, and compare each to the random baseline algorithm (RAND). To evaluate the impact of different preference orders on the performance of the IA, I compare each different version of this algorithm to IA-BASE. If ordering does make a difference, then any order determined by the data should perform quite well relative to GR and FB. On the other hand, comparison of various IAs to the relevant baseline order should show which orders are actually making a difference.

4.4 Evaluation Results

In what follows, pairwise comparisons between algorithms are reported using two-tailed t-tests by subjects (t_1) and items (t_2). Due to the large number of pairwise tests, all reported p-values are the result of a Bonferroni correction[9].

Pairwise algorithm comparisons should permit an answer to Q1 in §4.2.4, about the relative performance of different algorithms. Q2 is addressed by comparing the main IA preference orders in each dataset to the relevant IA-BASE. To assess the impact of different Cardinality/Similarity conditions (Q3), I report the results of univariate Analyses of Variance (ANOVA), taking both of these as independent variables. Once again, these tests are reported using participants (F_1) and items (F_2) as sources of variance.

Since the descriptions in the corpus were elicited under different conditions of MD, there may be variation in algorithm–human agreement as a function of the domain involved. On the other hand, an algorithm may well show good agreement with a subset of the participants (van Deemter, 2004). For this reason, I report means, modes (the most frequent score), and also the percentage of times an algorithm achieved perfect agreement with a human-authored description (a Dice score of 1), referred to as the *perfect recall percentage* (PRP).

4.4.1 Algorithm performance on the $-$LOC dataset

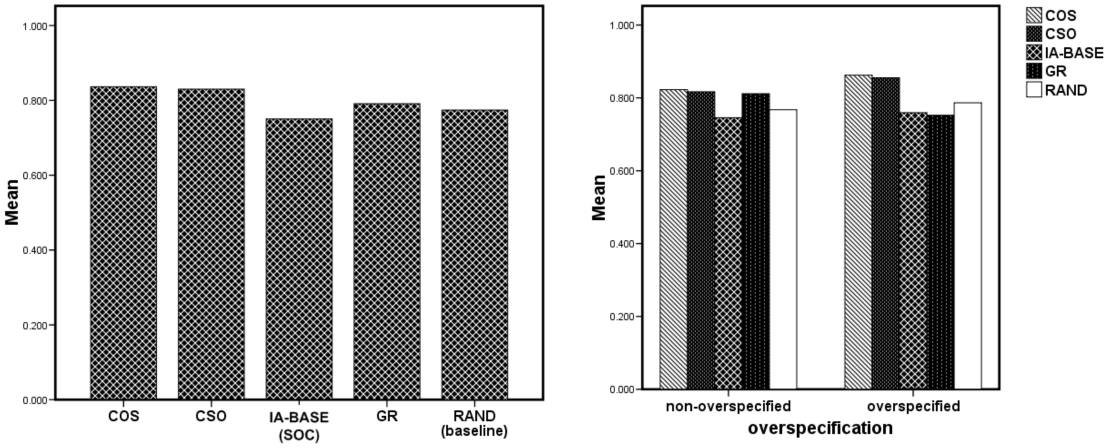
Figure 4.2(a) displays the mean Dice scores for the 3 algorithms tested on this dataset, as well as the RAND baseline and IA-BASE (the SOC order). Both COS and CSO perform marginally better than GR on this domain, but the difference does not appear enormous. IA-BASE performs worst of all. A slightly different perspective on the same data is offered by Figure 4.2(b), which shows agreement of the algorithms on those corpus descriptions where authors in the corpus had produced overspecified descriptions or not, according to the definitions given in §3.6.1 (p. 84). As expected, GR performs better on the non-overspecified descriptions, and the gap between GR and the two versions of the IA, COS and CSO, seems smaller on this data. However, the performance of GR declines on the overspecified instances, while the IA, though marginally better on these descriptions than on non-overspecified ones, declines much less dramatically. The figures show little if any difference between COS and CSO. This reflects the variance in the data in the relative preference of ORIENTATION and SIZE, arising from preferences for certain values of the former: the potential advantage of COS, which can include ORIENTATION when it is not strictly required, is cancelled out by its including it also when its values are dispreferred. Of course, this could be implemented in the IA using a preference order that ordered attribute-value pairs, rather than attributes.

[9] The correction is obtained by multiplying the $p-$value obtained on the test by the number of tests conducted, to avoid erroneous rejection of the null hypothesis.

−LOC			
	Mean	**Mode**	PRP
CSO	0.84	1	24.1
COS	0.83	1	24.1
GR	0.79	0.8	18.7
RAND	0.77	0.67	13.3
IA-BASE	0.75	0.67	7.4

+LOC			
	Mean	**Mode**	PRP
RAND	0.55	0.67	1.7
GR	0.58	0.67	5.8
FB	0.57	0.67	6.6
IA-BASE	0.54	0.67	1.7
CYSXO	0.66	0.67	10.2
CYSOX	0.64	0.67	10
CYOSX	0.64	0.67	10
COSYX	0.6	0.67	9.7
CSOYX	0.59	0.67	9.7
CSYOX	0.63	0.67	9.7
CSYXO	0.64	0.67	8.7
CYXSO	0.66	0.67	8.7
CYXOS	0.66	0.67	8.5
CYOXS	0.64	0.67	8.3
COYSX	0.62	0.67	7.5
YCXOS	0.65	0.67	6.1
YCXSO	0.65	0.67	6.1
COYXS	0.62	0.67	5.6
YCSXO	0.64	0.67	5.3
YCOXS	0.63	0.67	4.6
YXCOS	0.61	0.67	4.6
YXCSO	0.61	0.67	4.6
YCOSX	0.63	0.67	3.9
YCSOX	0.63	0.67	3.9

Table 4.1: Mean and modal scores of the algorithms in the two datasets



(a) Mean Dice score

(b) Dice score as a function of human overspecification

Figure 4.2: −LOC dataset: Mean Dice scores for each algorithm.

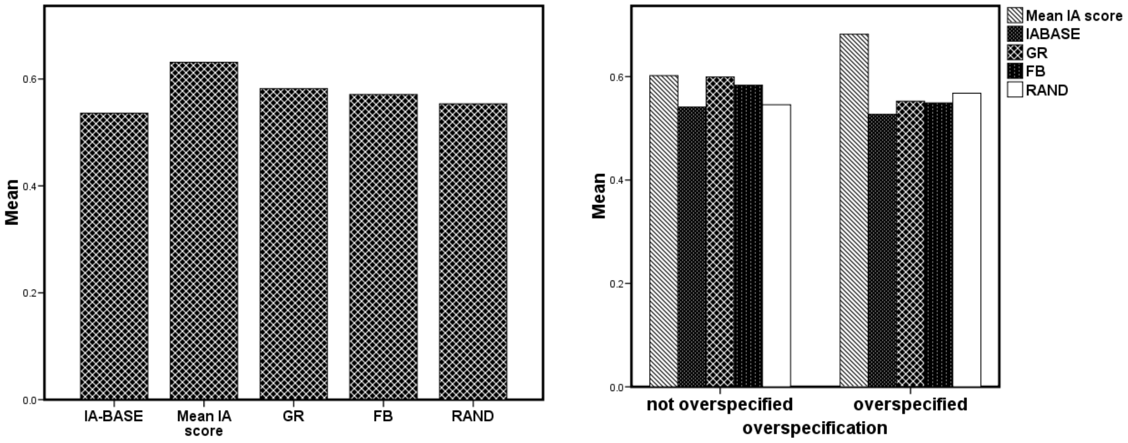
	GR	CSO	COS	IA-BASE
$t_1(26)$	3.333^*	9.620^*	7.002^*	-5.850^*
$t_2(19)$	1.169	5.241^*	4.632^*	-1.797

Table 4.2: −LOC: Pairwise comparison against the random baseline ($^*p \leq .01$)

The preliminary impressions offered by the figure are confirmed by the mean and modal values displayed in the top panel of Table 4.1, which also shows the PRP for each algorithm with a corpus description. Both IAs had a modal value of 1, achieved 24.1% of the time, while GR performed worse, achieving perfect agreement in 18.7% of cases. In this domain, RAND clearly fared worse than either IA and GR, while IA-BASE, the SOC order, performed worst of all, at least judging by the PRP it obtained. I shall return to this in §4.5. Despite its poorer performance, the 13% perfect score obtained by RAND is non-negligible. However, this dataset contains simple domains, and a random incremental procedure is more likely to converge on the same description as an algorithm and/or a human author a number of times, especially in those domains where MD contained SIZE, COLOUR, and ORIENTATION.

The simplicity of the domains tested here allows us to pose a version of the question raised by van Deemter (2004), namely, *What if an algorithm achieves* 100% *perfect agreement with a subset of individuals, or even only one?* None of the algorithms achieved this. The best match obtained with any single individual was a score of 1 on $35\% - 40\%$ of an individual's descriptions. Both COS and CSO achieved this 7 times, while GR achieved it 3 times.

Table 4.2 shows the results of pairwise comparisons between algorithms, including IA-BASE, to RAND. Only CSO and COS were significantly better both by subjects and items, though GR was better than RAND by subjects, and IA-BASE significantly worse. The lack of significance for GR by items reflects those cases where MD contained only dispreferred attributes. Here, authors tended to overspecify, especially with COLOUR; however, GR never includes these attributes in these cases, since they have little if any contrastive value. This results in significant variance over different



(a) Mean Dice score

(b) Dice score as a function of human overspecification

Figure 4.3: +LOC dataset: Mean Dice scores for each algorithm (a) overall and (b) distinguishing between overspecified and non-overspecified corpus descriptions

domains (items) in how well this algorithm performs.

A pairwise comparison of GR to COS and CSO showed that, although the better performance of the IA appears marginal in Figure 4.2, it was significant, both for CSO ($t_1(26) = 5.276$, $p < .001$; $t_2(19) = 2.526$, $p = .08$) and for COS ($t_1(26) = 2.972$, $p = .006$); $t_2(19) = 2.117$, $p = .08$), though only approaching significance by items after Bonferroni correction. Once again, the difference between the by-subjects and the by-items analysis, which shows that there was sizable variance from one domain to another, is due to the balanced setup of the corpus. Because domains represented all possible values of MD, there are instances where the dispreferred attributes are required to distinguish a referent. Though the proportion of underspecified descriptions was low, a small number of participants did tend to underspecify on precisely these domains, as shown in §3.8, increasing the variability of results from one domain to another.

Overall, the IA did perform better than a brevity-oriented procedure such as GR (or indeed FB, which is identical to GR on these domains), despite the simplicity of the domain. A better performance of all algorithms compared to RAND, with the exception of IA-BASE, also shows that even when content determination is limited to three attributes (and TYPE), the way people go about it is systematic, and does not involve a probabilistic sampling from the possible alternatives, although some early psycholinguistic models had suggested that this is indeed the process underlying referential communication (Rosenberg and Markham, 1971). A related result – perhaps the most crucial for an evaluation study of this nature – is the significant impact of preference orders on the IA, with the IA-BASE (SOC) version performing worse than RAND. Before considering the role of preference orders in more detail, let us turn to the results on the more complex dataset.

4.4.2 Algorithm performance on the +LOC dataset

Figure 4.3(a) displays the mean score obtained by the 20 versions of the IA, against those obtained by GR, FB, RAND and IA-BASE. Once again, the IA seems to have performed best overall, and

	$t_1(24)$	$t_2(19)$
FB	0.242	1.286
GR	0.544	1.900
COSYX	1.334	2.361
COYSX	2.561	3.498^{**}
COYXS	3.179^{**}	3.757^{*}
CSOYX	1.240	2.135
CSYOX	2.394	3.663^{**}
CSYXO	3.74^{*}	4.277^{*}
CYOSX	3.934^{*}	4.574^{*}
CYOXS	4.839^{*}	4.738^{*}
CYSOX	3.895^{*}	4.491^{*}
CYSXO	5.549^{*}	5.098^{*}
CYXOS	6.576^{*}	5.533^{*}
CYXSO	6.379^{*}	5.518^{*}
YCOSX	3.406^{**}	4.313^{*}
YCOXS	3.916^{*}	4.571^{*}
YCSOX	3.259^{**}	4.157^{*}
YCSXO	4.231^{*}	5.191^{*}
YCXOS	4.5^{*}	5.484^{*}
YCXSO	4.268^{*}	5.344^{*}
YXCOS	2.313	3.352^{**}
YXCSO	2.201	3.25^{**}
IA-BASE	0.705	1.776

Table 4.3: Pairwise comparisons to the random baseline, +Loc dataset. ($^{*}p \leq .02$; $^{**}p \leq .08$)

all the algorithms achieved a better score than the random baseline, except for IA-BASE. A comparison of the algorithms focusing on the difference between overspecified and non-overspecified descriptions (Figure 4.3(b)) reveals the same pattern as for the −LOC dataset: the IA is the better match to the human descriptions in both cases, although this time, there is a sharper rise in performance on the overspecified descriptions compared to the others. This is due to the fact that the mean averages over both those versions that placed the two preferred attributes – COLOUR and Y-DIMENSION – first, and all the others. Although this dataset contained more non-overspecified descriptions than the −LOC dataset, the corpus data suggests that placing these two attributes first would result in better performance. The bottom panel of Table 4.1, given above, confirms this. Though modal scores are constant for all the algorithms, and lower at .667 than in the previous dataset, the top three versions of the IA, which achieved perfect agreement roughly 10% of the time, are those placing COLOUR and Y-DIMENSION in that order, right at the top of the preference list. Note that switching the order of COLOUR and Y-DIMENSION makes performance decline. That the orders starting with YX perform the worst on this measure serves to confirm the earlier finding that the use of the X-DIMENSION attribute is relatively dispreferred, and an algorithm that uses COLOUR in place of this attribute fares better. Despite the evidence for preferences, however, there was significant variation among individuals in this dataset. None of the algorithms achieved perfect agreement with any individual: the best-performing versions of the IA agreed with an author at most 35% of the time. As with the other dataset, the worst-performing algorithm was the IA with baseline preference order.

Table 4.3 displays the results of pairwise comparisons of all the algorithms to the random baseline. The table indicates not only those means which were 'truly' significant after Bonferroni correction, but also some that approached significance at $p \leq .08$. Only some versions of the IA performed better than RAND. Like FB and GR, IA-BASE performed no better than RAND[10]. The reason for the much poorer performance of FB and GR on this dataset is that, although they do select locative attributes, they do not do so with the same frequency as those versions of the IA that rate these attributes highly. The chances of generating locative descriptions with a brevity-oriented strategy, using the inequalities algorithm of van Deemter (2006), depends on the extremity of the value. For instance, if a referent were the only entity in row 1 of a domain, it would be the sole entity in the extension of $\langle$Y-DIMENSION $< 2\rangle$, and this would make the vertical dimension highly discriminatory. For related reasons, GR and FB would presumably opt for the logically strongest inequalities whenever possible (they have higher discriminatory power), as does the IA because of the way gradables were ordered. However, briefer alternatives could often be found, resulting in descriptions which incurred mismatch to corpus instances, because they did not contain LOCATION.

Further support for these conclusions comes from the fact that the versions of the IA that exceeded the RAND baseline tended to be those with Y-DIMENSION towards the top of the preference order. Note, however, that the same pattern emerges here as with mean and modal values. Placing both X-DIMENSION and Y-DIMENSION first does not improve performance; it is the position of COLOUR and Y-DIMENSION with respect to other attributes in the preference order that determines the overall performance of the algorithm. This echoes the result of §3.8.1 (p. 91), which showed that, despite a reliable decrease in the use of 'inherent' perceptual properties when perspective shifted to the 'where' of an entity, the use of COLOUR still tended to be higher than that of other attributes. This is also related to a recent finding by Paraboni et al. (2006), who showed that people's resolution of locative expressions in hierarchical domains could be facilitated by overspecification.

A comparison of GR to FB revealed that the small difference in their mean scores (see the bottom panel of Table 4.1), was not significant ($t_1(24) = .773$, ns; $t_2(19) = 1.455$, ns). This does not mean that they gave identical output in all cases. For example, in a domain where MD contained COLOUR, ORIENTATION and SIZE, both FB and GR generated a briefer description by considering LOCATION. One of the human-authored descriptions in this domain is shown in (4.8)

(4.8) $\langle$TYPE : *chair*$\rangle \wedge \langle$Y-DIMENSION $=$ *1*$\rangle \wedge \langle$COLOUR : *grey*$\rangle \wedge \langle$X-DIMENSION $=$ *4*$\rangle$

In this domain, only GR included COLOUR, as shown in (4.9), while FB didn't (4.10) including X-DIMENSION in its stead.

(4.9) $\langle$TYPE : *chair*$\rangle \wedge \langle$COLOUR : *grey*$\rangle \wedge \langle$X-DIMENSION $>$ *3*$\rangle$

(4.10) $\langle$TYPE : *chair*$\rangle \wedge \langle$Y-DIMENSION $<$ *2*$\rangle \wedge \langle$X-DIMENSION $=$ *4*$\rangle$

[10] The $p-$value obtained for GR was .04, but this fails to reach significance after Bonferroni correction

	FB		GR	
	$t_1(24)$	$t_2(19)$	$t_1(24)$	$t_2(19)$
COSYX	-1.318	-0.836	-1.148	-0.359
COYSX	-2.479	-1.523	-2.300	-1.073
COYXS	-2.944	-1.645	-2.688	-1.180
CSOYX	-1.209	-0.708	-1.037	-0.262
CSYOX	-2.835	-2.087	-2.824	-1.609
CSYXO	-4.639^*	-2.816	-4.5^*	-2.357
CYOSX	-4.235^*	-2.539	-4.092^*	-2.091
CYOXS	-4.981^*	-2.656	-4.613^*	-2.201
CYSOX	-4.149^*	-2.517	-4.067^*	-2.079
CYSXO	-5.52^*	-3.215	-5.018^*	-2.806
CYXOS	-5.966^*	-3.478	-5.145^*	-3.055
CYXSO	-5.72^*	-3.464	-5.024^*	-3.044
YCOSX	-3.845^*	-2.248	-3.072	-1.723
YCOXS	-4.084^*	-2.442	-3.127	-1.898
YCSOX	-3.638^*	-2.172	-2.956	-1.664
YCSXO	-4.306^*	-3.032	-3.386^*	-2.501
YCXOS	-4.2^*	-3.256	-3.31^*	-2.717
YCXSO	-3.978^*	-3.171	-3.172^*	-2.639
YXCOS	-1.700	-1.315	-1.296	-0.839
YXCSO	-1.604	-1.261	-1.215	-0.788

Table 4.4: +LOC: IA versus GR and FB ($^*p \leq .02$)

In contrast to these two, the CYSXO version of the IA, which was one of the better-performing orders, matched the human description perfectly on this domain, because SIZE was not distinguishing, so that the final description contained the same attributes as the human one because the algorithm included COLOUR, Y-DIMENSION, and X-DIMENSION. The better versions of the IA were more consistent than FB and GR in including preferred attributes, since this did not depend exclusively on discriminatory power or brevity.

Pairwise contrasts between each version of the IA, and FB and GR, are shown in Table 4.4. The figures reflect the picture presented by the modes and mean scores in Table 4.1. It is those orders which place COLOUR and Y-DIMENSION first which perform significantly better than either GR or FB. The others are not significantly different.

To summarise, this dataset, like the previous one, shows a superiority of the IA, but this is only the case with some preference orders that reflect human preferences in the corpus. One general conclusion that can be reached is that a strategy that aims to achieve brevity, or approximate it, will not match human preferences. This is even more the case as the domain gets more complex, and remains true despite the fact that TYPE is always included. The evaluation results have something to say about what people do, lending further support to the results of the data analysis, showing that there are highly systematic preferences operative on how people perceive, conceptualise and describe objects. From this point of view, the apparent dependency of the performance of the IA on its preference order only serves to strengthen the interpretation of the *gestalts* hypothesis given in §3.8 (p. 89). To make this conclusion more precise, let us now turn to the comparison of the

	$t_1(24)$	$t_2(19)$
CSYXO	-3.269^*	-6.193^*
CYOSX	-3.300^*	-6.22^*
CYOXS	-3.934^*	-6.928^*
CYSOX	-3.265^*	-6.22^*
CYSXO	-4.642^*	-7.059^*
CYXOS	-5.678^*	-8.026^*
CYXSO	-5.566^*	-8.104^*
YCOXS	-3.252^*	-7.412^*
YCSXO	-3.72^*	-8.008^*
YCXOS	-4.239^*	-9.262^*
YCXSO	-4.03^*	-9.023^*
YXCOS	-2.777	-8.101^*
YXCSO	-2.644	-7.844^*
COSYX	-1.408	-3.525^*
COYSX	-2.43	-4.602^*
COYXS	-2.962	-5.286^*
CSOYX	-1.322	-3.378^*
CSYOX	-2.18	-5.406^*
YCOSX	-2.797	-6.426^*
YCSOX	-2.688	-6.300^*

Table 4.5: +LOC: Comparison of different versions of the IA to IA-BASE ($^*p \leq .02$)

various IAs to their baseline orders.

4.5 The impact of preference orders on the Incremental Algorithm

In §2.7, I argued that there are as many versions of the IA in a given domain as there are possible preference orders. This is both a strength and a weakness. On the one hand, it makes it possible for the algorithm to reflect both general and domain-specific preferences; on the other, it makes the algorithm difficult to falsify, unless a domain can be found in which all preference orders perform equally well. This was certainly not the case in the present study, where IA-BASE in the two datasets clearly performed worse even than GR, FB and RAND.

In the $-$LOC dataset, IA-BASE (SOC) performed much worse than COS ($t_1(26) = 10.725$, $p < .001$; $t_2(19) = 4.112$, $p = .001$) and CSO ($t_1(26) = 13.065$, $p < .001$; $t_2(19) = 3.829$, $p = .001$). As shown in Table 4.1, CSO and COS were indistinguishable on the basis of their degree of match to human data. Results of the pairwise comparisons on the more complex +LOC dataset are given in Table 4.5. All the orders performed significantly better than IA-BASE by items, when domains are considered as the source of variance. This reflects the pattern of results reported above: while IA-BASE in +LOC was significantly worse than RAND, none of the 'real' orders were, though some were indistinguishable from the random algorithm. Many of the orders that performed no better than RAND do not emerge as significantly better than IA-BASE once authors are the source of variation, as the by-subjects figures in the table (t_1) indicate.

In §3.8.1 (p. 91), some evidence was cited for the idea that authors select a 'perspective' on a referent. The analysis of frequencies of usage of attributes showed that when LOCATION

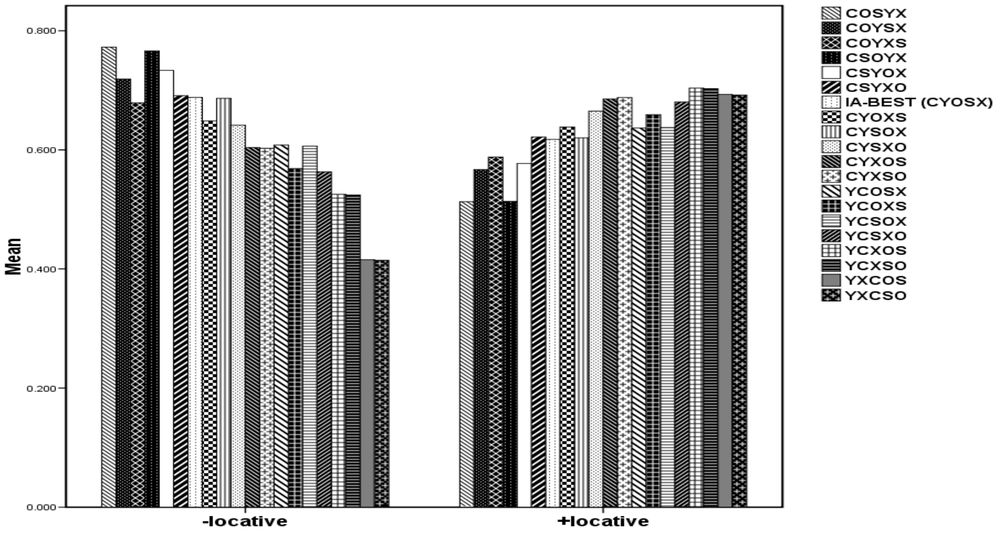


Figure 4.4: Performance of different preference orders as a function of locative use

(in either dimension) was used by an author, there was a corresponding tendency to use the other attributes less. The trends in the data still conformed to the basic preferences observed in the analysis of usage of attributes; in particular, the use of SIZE and ORIENTATION declined with the use of LOCATION, more than did the use of COLOUR. Yet, overall, the data did suggest a shift in perspective from the 'what' to the 'where' of an entity. Although the two perspectives are clearly not mutually exclusive (otherwise, attributes other than LOCATION would never be used in a locative description), they are bound to have an impact on the performance of the IA. The data in the +LOC dataset offers the possibility of investigating this. Recall that this dataset contained descriptions from authors who had used LOCATION in conditions where they were instructed not to, as well as data from those authors in the +FC+LOC condition. Among the latter, the extent to which an author used locatives in their descriptions varied, so that not all authors used the 'where' perspective 100% of the time.

The preference orders for the IA on this dataset were obtained by interpolating X- and Y-DIMENSION with the two preferred orders for the other attributes, namely COS and CSO. Viewing the performance (as indicated by the mean Dice score) of the different IAs as a function of whether a description contained a locative expression may shed some further light on the question of the impact of preference orders. Figure 4.4 displays the mean score obtained by each of the 20 algorithms, as a function of whether descriptions contained locatives.

The most interesting aspect of the trends in the Figure is that the best-performing versions of the IA on one set of descriptions become the worst-performing on the other. This is most dramatic with versions whose means and modes are displayed in Table 4.6.

Two orders – COSYX and CSOYX – are by far the best-performing versions on descriptions which had no locatives, but emerge as the worst on locative descriptions, with a corresponding decline in mean score. Similarly, YCXOS and YCXSO exceed IA-BEST, the order found to be the best-performing overall (CYOSX). This is because they order Y-DIMENSION before COLOUR, but

	Locative		No Locative	
	Mean	**Mode**	**Mean**	**Mode**
CYOSX (IA-BEST)	.688	.667	.618	.667
COSYX	.772	1	.513	.667
CSOYX	.766	1	.514	.667
YCXOS	.525	.667	.704	.667
YCXSO	.524	.667	.703	.667

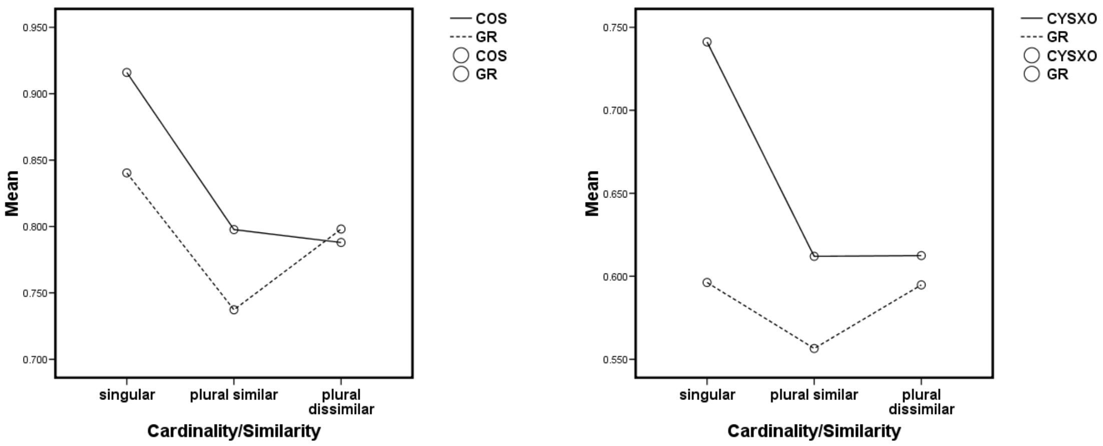
Table 4.6: Means and modes for the best- and worst-performing algorithms, by LOCATION

are still likely to use the latter because it is highly preferred (mirroring the lower decline in the use of this attribute in the human data when LOCATION was included in a description). Moreover, they also have a higher tendency to select X-DIMENSION, since this is ordered just after COLOUR. Thus, this attribute is often used with Y-DIMENSION, which is what the data analysis in §3.8.1 would predict. For the same reasons, performance declines on the non-locative descriptions, which is where COSYX and CSOYX perform much better. Against this background, the best overall performance of the CYOSX order – also indicated by its relatively constant mean score on descriptions with or without location – suggests that this is a compromise solution between two trends that pull in opposite directions.

These remarks are meant to highlight a further feature of referring expressions generation, when this is compared against human performance. The latter can be unpredictable when people are not self-consistent in their use of specific attributes, as happens here when not everybody consistently used LOCATION in their descriptions. Related issues have been raised by Reiter and Sripada (2002a), who observe that weather forecasters are often inconsistent in their choice of words or phrases to express the same temporal interval. In relation to perspective-taking, a statistical analysis can suggest that certain trends hold in the data, and the evaluation results confirm it. However, because these trends tend not to be all-or-none, the performance of an algorithm can vary as a function of individual variation.

These results show that whether a domain is simple or complex, a very significant impact of the preference order can be observed. This has implications for the interpretation of the results of the evaluation. Does the evaluation show that the IA is superior to GR and FB? I believe that a precise reply to this question is not possible. It is only possible to say that there exist versions of the IA which perform better. The general features of the IA that make it a better match to the human data are (a) its slight tendency to overspecify (cf. Figures 4.2(b) and 4.3(b)), which is what humans do, and (b) its tendency to build descriptions incrementally along the gradient represented by its order. Other than that, the features of specific incarnations of this algorithm will determine its performance.

These conclusions raise some questions of broader relevance to NLG: If an algorithm is so strongly bound to hand-coded preferences, are results concerning performance of the algorithm ever generalisable? I think the answer to this question is only positive if the preferences found have some generalisability beyond the domain in which they are tested. In the present case, one could argue that at least some of the preferences identified are ‘general’, in the sense that they



(a) −LOC dataset

(b) +LOC dataset

Figure 4.5: Performance of the algorithms as a function of Cardinality/Similarity

are supported by previous psycholinguistic research. However, even preferences such as those identified in Chapter 3 could turn out to be violable, for example because of cross-genre variation. For example, a particular variety of language could impose explicit constraints on how objects are described because of a highly fault-critical communicative setting. In such cases, convention might be a stronger determining factor in content determination than conceptual or perceptual primacy. What of domains in which such preferences are unknown, or are impossible to discover (for instance, domains where no subset of the set of possible versions of the IA clearly outperforms all others)? Is it feasible to require of the NLG system builder an exhaustive empirical analysis of domain-specific corpora, or is it sufficient to formalise preferences in a simpler fashion, for example by making frequency counts of attribute selections in a corpus of the right genre? The procedure recommended by Reiter and Dale (2000) is a corpus analysis that serves as a pilot study, whose outcome informs the design of hand-coded rules. Another option, which has only recently begun to be exploited in NLG work, is to use corpus-based language models to act as 'filters' for the output of a system. This idea seems to work quite well in surface realisation, where multiple outputs of a grammar can be evaluated against data, to select the best candidate (e.g. Langkilde, 2000). So far, the only proposal of this nature in GRE has been Varges (2004) (on which see §2.7.7, p. 63); however, this methodology requires a semantically annotated corpus to be viable, given that a core part of GRE is content determination. In general, the main practical issue is the availability of adequate resources to inform content determination.

4.6 Singulars and plurals

The final part of this analysis focuses on the difference in performance of the algorithms when generating references to singletons versus plurals. The variable of interest here is Cardinality/Similarity. Since the previous sections established that there was no difference between GR and FB[11], only GR and IA are compared in this section. Moreover, to avoid a plethora of statistical tests, I focus on only one version of the IA in each dataset, namely COS for −LOC and CYSXO for

−LOC. The performance of the IA and GR in the two datasets is displayed in Figure 4.5.

One of the salient features in these figures is the decline in performance of the IA in both datasets when it generates plurals, that is, when the version of the IA is van Deemter's (2002) IA$_{bool}$. There is little difference between the Plural Similar and the Plural Dissimilar cases. The performance of GR, on the other hand, while better overall on singletons, improves slightly on the Plural Dissimilar domains in the −LOC dataset. Interestingly, its performance on these domains equals that on Singulars in the +LOC dataset, while there is still a decline on Plural Similar domains, though the drop is slight.

Separate ANOVAs with Cardinality/Similarity as independent variable were conducted for the two datasets, for each algorithm. In the −LOC dataset, the difference between conditions was significant both for COS ($F_1(2, 23) = 50.367$, $p < .001$; $F_2(2, 17) = 40.095$, $p < .001$), and for GR, though in the latter case, the effect was not reliable by items ($F_1(2, 23) = 22.1$, $p < .001$; $F_2(2, 17) = 2.171$, ns). Pairwise comparisons between different levels of Cardinality/Similarity were carried out, using a Bonferroni test to estimate truly significant differences at $\alpha = .05$. In the case of COS, this showed that the only significant differences were those between the Singular case, and the two Plural cases. There was no difference between Plural Similar and Dissimilar domains, precisely as Figure 4.5(a) leads one to expect. For GR, all pairwise differences turned out significant, confirming its apparent improvement on the Dissimilar cases compared to the Similar ones.

The same calculations on the +LOC dataset yielded much the same results for the IA, with a significant main effect ($F_1(2, 21) = 17.024$, $p < .001$; $F_2(2, 17) = 10.275$, $p = .001$), and identical results for pairwise contrasts. There was no significant effect for GR, implying that it performed equally on singletons and plurals. The mean difference in the Dice coefficient between any pair of conditions for this algorithm in fact never exceeded .5, as a glance at Figure 4.5(b) will confirm.

Overall, however, it is reasonable to conclude that the algorithms performed worse when they generated references to multiple entities, whether or not the minimal description required for those entities was a conjunction (the Plural Similar case) or a disjunction (Plural Dissimilar).

Consider an example of the output of IA and GR in the Plural Similar condition, shown in (4.11). Because TYPE is included by default by both algorithms, and the values of TYPE differ for plural referents in this, as in the Plural Dissimilar Condition, both algorithms produce a disjunction of two TYPE values. However, the rest of the description is a conjunction of properties, because neither of the algorithms needs to consider disjunctions before finding a distinguishing description.

(4.11) (a) GR:

$$(\langle \text{TYPE} : \textit{desk}\rangle \vee \langle \text{TYPE} : \textit{fan}\rangle)$$
$$\wedge$$
$$\langle \text{Y-DIMENSION} > \textit{1}\rangle \wedge \langle \text{X-DIMENSION} > \textit{3}\rangle \wedge \langle \text{ORIENTATION} : \textit{front}\rangle$$

(b) IA:

$$(\langle \text{TYPE} : \textit{desk}\rangle \vee \langle \text{TYPE} : \textit{fan}\rangle)$$
$$\wedge$$
$$\langle \text{COLOUR} : \textit{red}\rangle \wedge \langle \text{SIZE} : \textit{large}\rangle \wedge \langle \text{ORIENTATION} : \textit{front}\rangle$$

[11] That is, FB and GR return identical output on the −LOC dataset, while no significant difference in performance was found between them on the +LOC dataset.

An example of a human-produced description in the same domain is shown in (4.12). There is a high degree of redundancy because identical properties are repeated to describe elements, and the description corresponds to a partition of the set. The automatically-produced descriptions in the same domain, while differing in content in the case of GR, differ also in their structure. Note that the IA produces essentially the same content as the human author, while GR doesn't. This is a consequence of the tendency to overspecify as a result of the gradient descent strategy, and is why IA outperforms GR most of the time in different Cardinality/Similarity conditions. However, both algorithms are penalised here for having fewer occurrences of the properties $\langle \text{COLOUR} : \textit{red} \rangle$, $\langle \text{ORIENTATION} : \textit{front} \rangle$ and $\langle \text{SIZE} : \textit{large} \rangle$ than the human-authored description.

(4.12) HUMAN:

$$(\langle \text{SIZE} : \textit{large} \rangle \wedge \langle \text{TYPE} : \textit{desk} \rangle \wedge \langle \text{ORIENTATION} : \textit{front} \rangle \wedge \langle \text{COLOUR} : \textit{red} \rangle)$$
$$\vee$$
$$(\langle \text{SIZE} : \textit{large} \rangle \wedge \langle \text{TYPE} : \textit{fan} \rangle \wedge \langle \text{ORIENTATION} : \textit{front} \rangle \wedge \langle \text{COLOUR} : \textit{red} \rangle)$$

Is it therefore possible to conclude that redundancy is the norm in human descriptions? The data suggests that it is not informativeness that is playing the crucial role here, but the way in which people conceptualise and describe sets by partitioning when entities belong to disjoint types. This will be the central topic of the forthcoming chapter, where a different algorithm for the generation of references to sets is tested.

It could be argued that the algorithms, especially the IA, are being unfairly penalised in these instances since, for example, (4.11b) is logically equivalent to (4.12). However, I opted to leave the logical forms intact, without normalisation, in order to explore the strategy that algorithms such as IA$_{bool}$ incorporate, relative to human strategies. Despite their logical equivalence, the descriptions are transparent reflections of different generation/production strategies, one which proceeds by describing salient elements of a set (perhaps comparing them to each other – cf. §3.9, p. 94), and the other by conjoining disjunctions of increasing length in a more 'naively incremental' fashion.

This brings up another question, namely why GR fares better on Plural Dissimilar domains than Similar ones, at least in the $-$LOC dataset while IA fares slightly (though not significantly) worse. This is because in this case, GR avoids redundancy – precisely the pattern observed in the corpus data with Dissimilar plurals, where a larger proportion of well-specified descriptions were found compared to the Singular and Plural Similar case (see §3.9, p. 94). There was some evidence that authors adopted a contrastive strategy, in that having two salient referents in focus, differing on the critical dimensions, facilitated the process of content selection of those dimensions, thereby reducing the tendency to overspecify. The IA in these cases is more redundancy-prone. Perhaps more seriously, it tends to produce disjunctive descriptions that involve overlaps between disjuncts. For instance, in (4.13), generated by the Incremental procedure, Y-DIMENSION occurs twice, with one occurrence in the disjunction with *grey*.

(4.13)

$$(\langle \text{TYPE} : \textit{desk} \rangle \vee \langle \text{TYPE} : \textit{sofa} \rangle)$$
$$\wedge$$
$$(\langle \text{X-DIMENSION} < \textit{4} \rangle \wedge \langle \text{Y-DIMENSION} > \textit{1} \rangle \wedge \langle \text{ORIENTATION} : \textit{right} \rangle)$$
$$\wedge$$
$$(\langle \text{COLOUR} : \textit{grey} \rangle \vee \langle \text{Y-DIMENSION} > \textit{2} \rangle)$$

Note that the first occurrence of Y-DIMENSION is true of both referents and, though the second is the logically stronger inequality, it is included after the first (which is true of both referents in the domain), because it is part of a disjunction, and the IA considers disjunctions after literals. In this case, applying the inference rules for gradables did not change the picture, because $\langle$Y-DIMENSION $> 2\rangle$ is not true of both referents, and removing it would result in an unsuccessful description. The description contains a disjunction that divides the set of referents into two sets: 'things which are grey' and 'things which are below the middle row'. However, it also predicates the property 'below the top row' of both sets. In this sense, the description contains an overlap between those elements of which the location property is predicated, and those of which the colour property is predicated: It turns out that there is a location property that applies to both. This may be avoided by human authors because a description of the form $\phi \vee \psi$ carries the implicature that $[\![\, \phi \,]\!] \cap [\![\, \psi \,]\!] = \emptyset$.

To summarise, the performance of the algorithms on plurals brings up two classes of issues. The first has to do with logical transparency. IA$_{bool}$ and the Boolean version of GR often produce opaque descriptions, whose mapping to an NL representation is unclear. Considerable simplification of the logical forms, perhaps along the lines suggested by van Deemter (2002), would often have to be carried out. The second, related, issue is the strategy that a generator should take in describing a set. Examples such as (4.12) suggest that humans *partition* sets whose elements have different basic-level TYPE values. This might indicate that a partitioning algorithm, such as that proposed by van Deemter and Krahmer (2006), might do the trick. Nevertheless, the overspecification data presented in the previous chapter, as well as the strong evidence for attribute preferences, suggest that people do not simply search for an arbitrary partition whose elements can be described non-disjunctively (which is what the van Deemter and Krahmer algorithm does).

These questions will feature heavily in the chapters to follow. In line with the methodology of this thesis, I propose to first ask what it is that people do when they describe a set, and how their descriptive strategies depend on the domain, and the similarity of the referents. From there, I will move on to algorithms. Although the starting point in the next chapter is the IA, the aims of subsequent chapters are to find more general principles underlying the conceptualisation and incremental description of sets.

4.7 Summary and outlook

This chapter built on the previous one, taking as a starting point some of the results of the corpus analysis and using them to inform an evaluation study of the GRE algorithms that characterise the state of the art. Although some previous studies have been done in the area, this was the first GRE evaluation on this scale, which moreover extended coverage to include plurality and numeric-valued attributes. As shown in Chapter 2, algorithms that extend coverage in this manner can be separated from the search strategy proper that the IA, FB and GR algorithms incorporate. Hence, each of these strategies can be extended to deal with logically more complex referring expressions, and even gradable properties.

The conclusions of this evaluation can be summarised as follows. The Incremental Algorithm outperformed its predecessors, but this result should be discussed in the light of the very clear dependency of this algorithm on a hand-coded parameter. In conjunction with the data analysis, the better performance of the good versions of the IA can be viewed as a falsification of the hypotheses about referential adequacy that brevity-oriented strategies incorporate.

The results also highlight an important difference between the three algorithms tested here. It is relatively straightforward to give a formal, declarative statement of what GR and FB achieve, and to predict what their output will be, given the input. This is not in general possible for the IA, where its better performance in relation to the other two algorithms was shown to be highly dependent on the predefined preference order.

Like the earlier corpus analysis, this evaluation also tackled the issue of plurality from an empirical point of view for the first time. Plurality raises particularly difficult issues, in that there are questions related both to the ideal strategy for describing sets, and to the extent to which those strategies should be linguistically constrained, as some authors have suggested in the past (e.g. Horacek, 2004), to make the generation of logical forms as linguistically transparent as possible. To date, there has been no empirical work to back proposals made in the literature in this area. The next few chapters seek to build on the empirical groundwork laid in this and the previous chapter. The point of departure will be a comparison of IA$_{bool}$, and a new partitioning algorithm that is based on the corpus data. A more in-depth analysis of plurals in the corpus will also allow the formulation of some principles whereby people describe sets, which will be extended in later chapters and lead to further revisions of the basic generation strategy.

Chapter 5

Sets, gestalts, and partitioning

5.1 Introduction

This chapter marks the beginning of the second part of this thesis, where the focus is exclusively on plural reference. So far, some evidence has been gathered from the TUNA Corpus regarding attribute preferences and the tendency to overspecify not only in singular but also in plural descriptions. In addition, an evaluation of classic GRE algorithms has shown that although some algorithms perform well when compared to human data, plurality remains a problem.

This chapter extends the analysis of the plural data in the corpus described in Chapter 3, and proposes a new Content Determination algorithm, generalised to deal with plurals. In Chapter 4, a comparison of some examples of plural descriptions from the corpus to those returned by the Incremental (IA) and Greedy (GR) algorithms in the same domains suggested that the decline in performance on plurals was due to a mismatch in descriptive strategies between humans and algorithms, when the latter are extended to deal with disjunction in the manner proposed by van Deemter (2002). Although van Deemter's algorithm guarantees full Boolean completeness for GRE (see the discussion in §2.7.5, p. 58), it represents a primarily *logical* take on the problem, constructing a distinguishing description in Conjunctive Normal Form. The result differs in interesting ways from the initial examples from the corpus, reproduced below from §4.6 (p. 123).

(5.1) IA:

$$(\langle \textsc{type} : \textit{desk}\rangle \vee \langle \textsc{type} : \textit{fan}\rangle) \wedge \big((\langle \textsc{colour} : \textit{red}\rangle \wedge \langle \textsc{size} : \textit{large}\rangle \wedge \langle \textsc{orientation} : \textit{front}\rangle\big)$$

(5.2) HUMAN:

$$\begin{array}{c} (\langle \textsc{size} : \textit{large}\rangle \wedge \langle \textsc{type} : \textit{desk}\rangle \wedge \langle \textsc{orientation} : \textit{front}\rangle \wedge \langle \textsc{colour} : \textit{red}\rangle) \\ \vee \\ (\langle \textsc{size} : \textit{large}\rangle \wedge \langle \textsc{type} : \textit{fan}\rangle \wedge \langle \textsc{orientation} : \textit{front}\rangle \wedge \langle \textsc{colour} : \textit{red}\rangle) \end{array}$$

These two descriptions are logically equivalent, but there are some salient aspects of the human description (5.2) that are worth investigating further. Though the different aspects are highly related, I list them separately here, together with pointers to the sections below in which they will be analysed.

1. The human description represents a partition of the set, consisting of two disjuncts, each of which describes an entity. As the analysis in §5.2.1 (p. 131) will show, such *partitioned descriptions* are the norm in the TUNA Corpus. However, there is also an effect of Similarity,

insofar as fewer partitions are evinced in the Similar compared to the Dissimilar experimental condition.[1] The explanation I will offer for the general tendency to partition centres on the fact that elements of a target referent set in TUNA had different basic-level TYPE values. The data strongly suggests that humans begin by categorising the objects and then adding further information to the resulting (disjoint) segments of their descriptions. This finding is at the basis of the first principle proposed below, which states that reference is *category-driven*.

2. The description does not contain *aggregation*, that is, the author of (5.2) did not choose to write *the large red desk and fan facing front*, in which (the realisation of) the properties $\langle$SIZE : *large*$\rangle$, $\langle$COLOUR : *red*$\rangle$ and $\langle$ORIENTATION : *front*$\rangle$ modify both $\langle$TYPE : *desk*$\rangle$ and $\langle$TYPE : *fan*$\rangle$. As a result of this, the description contains a significant degree of redundancy. To borrow a term from Gardent (2002), this is a kind of *epistemic* redundancy, whereby the same properties are predicated of two objects, and propagated across two disjuncts in the description.[2] IA_{bool} avoids this because it first attempts to conjoin literals to the description, which in this case results in success (cf. 5.1 above). The analysis in §5.2.1 (p. 131) will also show that aggregation was rare overall.

3. Because of the partitioning and propagation strategy, the description exhibits a significant amount of *semantic parallelism*, where each element of a partition is described using exactly the same attributes. As shown in §5.2.2 (p. 134), this is also a feature of most descriptions in the corpus. The interpretation I offer of this trend is related to the Gestalts Hypothesis of Pechmann (1989) and the results on codability/attribute preferences obtained in Chapter 3. In particular, the analysis shows that the likelihood of an attribute being propagated across elements of a partitioned description is directly predictable from its codability, which one indicator of which is the likelihood of usage of such an attribute when it is not required for a distinguishing description. Thus, COLOUR is likely to be repeated in a description such as *the blue chair and the blue sofa*, whereas propagation of less preferred attributes like SIZE is less likely. In addition, the propagation of an attribute increases the similarity in the ways elements of a partition are conceptualised. A description such as *the blue sofa and the blue chair*, produced in the Plural Similar condition, draws a hearer's attention to attributes that make elements of a set similar and therefore, by hypothesis, more easily perceived as a group. This harks back to Wertheimer's principles of group perception (Wertheimer, 1938), where similarity indeed plays a central role. It is enforced by a further observation: parallelism is very strongly in evidence also in the Plural Dissimilar condition[3], but there is significantly more of it in the Similar condition.

Following the data analysis, §5.3 (p. 140) describes the design of the algorithm, which is presented here as an extension to the Incremental Algorithm (IA), but will also serve as the basis

[1] To recapitulate: The Similar condition consisted of two referents which differed on their TYPE but were otherwise identical; in the Dissimilar condition, the referents differed on all their attributes.

[2] The term 'propagation' refers to the fact that properties such as $\langle$COLOUR : *red*$\rangle$ in this example are used in two disjoint parts of the description.

[3] In this condition, an analogous description to the one exemplified for the Similar case might be *the blue sofa and the* red *chair*.

for work in Chapter 7, which moves beyond it. The characteristics of the algorithm, based on the data analysis, can be summarised as follows:

1. **Structure**: The algorithm presented here aims to produce logical forms whose structure is as close as possible to the surface form that the corpus data suggests is the norm among authors. This is achieved by initially partitioning a set by the basic-level TYPE of its elements.

2. **Content and similarity**: The algorithm incorporates a corpus-derived statistical model to predict when an attribute should be propagated, modulo its codability. The latter is operationalised in probabilistic terms, as the likelihood of propagation of an attribute even when it is not required. Therefore, this algorithm may sometimes include a property not because it has discriminatory value, but because it will enhance the semantic parallelism in the description.

3. **Strategy**: In order to generate descriptions in the manner indicated above, the algorithm does not search through disjunctions of increasing length, selecting those combinations which are true of the set of intended referents and have some contrastive value. Rather, it uses the information in the Knowledge Base to partition a set of intended referents opportunistically. Every time a property is true of *some* intended referents, the set is partitioned, and this is reflected in the structure of the description. This strategy may result in further partitioning after the initial category-driven partitioning step, returning partitioned descriptions of sets whose elements are of the same TYPE but have different values of other attributes (e.g. *the blue chair and the red chair* rather than *the blue and red chairs*). Since the TUNA Corpus does not contain plural references to entities of the same TYPE, this potential shortcoming is addressed later, in §5.5 (p. 154), where a new corpus study is conducted, the semantic investigating semantic constraints on aggregation (wide-scope modification) within plural NPs, and their syntactic complexity limitations. The original algorithm is then extended with a new procedure to deal with aggregation.

4. **Efficiency**: Since partitioning is opportunistic, and search is only carried out through literals, the algorithm has polynomial complexity.

This algorithm is evaluated against the remainder of the corpus data in §5.4 (p. 150), where it is shown to outperform IA_{bool}.

5.2 Data analysis

Because this chapter focuses in more depth on plurals, and applies conclusions from the corpus data to algorithm design, the analysis uses only a subset of the data, reserving the remainder for evaluation. The corpus contains 585 plural descriptions, from 45 authors, each of whom produced 7 in the Plural Similar condition, and 6 in the Plural Dissimilar condition. From this, a stratified random sample of 180 descriptions, referred to as PL_1, was generated by randomly taking 2 descriptions from the Similar and Dissimilar conditions from each author in the corpus, leaving a separate dataset (PL_2; $N = 405$) to be used for the evaluation reported later. Thus, the sample PL_1 contained an equal number of representatives from each author and an equal number of plural references elicited in the Similar and Dissimilar conditions. The difference between these two conditions is summarised as follows (see also §3.9, p. 94):

1. In the Similar condition, referents have different values of TYPE, but identical values on all attributes, whether these attributes were contrastive or not. Example (5.2) was elicited in this condition.

2. In the Dissimilar condition, referents have different values of TYPE *and* on the contrastive attributes, but identical values on all other attributes.

.

The differences between the two conditions do not include location, which was always randomly determined. Authors in the Similar condition had a variety of options to describe a set. For instance, the author who produced (5.2) could equally well have written either of the following.

(5.3) (a) the large red desk and fan facing front

(b) the large red (furniture items/objects) facing front

I refer to (5.3a) as an **aggregated, disjunctive** description, in that the realisations of $\langle$SIZE : *large*$\rangle$ and $\langle$ORIENTATION : *right*$\rangle$ have wide scope over the coordinate NP *desk and fan* (which is logically a disjunction). By contrast, example (5.2) above is **non-aggregated**, or **partitioned**, and contains considerably more redundancy. Example (5.3b) corresponds to a logical conjunction; this is made possible by the use of a superordinate term such as *furniture items* or *objects* for the referents. Though the example given so far is based on a Plural Similar domain, descriptions in the Plural Dissimilar condition can also be aggregated if they contain redundant information. This is because the referents have identical values on the non-distinguishing attributes. For example, suppose two referents had the attributes shown in (5.4) and the minimally distinguishing description consisted of COLOUR. A possible overspecified description of the referents is *the small, front-facing blue fan and green sofa*, giving ORIENTATION and SIZE wide syntactic scope over the NP *blue fan and green sofa*. Moreover, where LOCATION is used in either condition, there is always the possibility of referring using a non-disjunctive description (e.g. *the objects in the top row*).

(5.4) (a) $\Big\{\langle$ TYPE: *fan* $\rangle$, $\langle$ ORIENTATION: *front,* $\rangle$, $\langle$ COLOUR: *blue* $\rangle$, $\langle$ SIZE: *small* $\rangle\Big\}$

(b) $\Big\{\langle$ TYPE: *sofa* $\rangle$, $\langle$ ORIENTATION: *front,* $\rangle$, $\langle$ COLOUR: *green* $\rangle$, $\langle$ SIZE: *small* $\rangle\Big\}$

The data analysis reported below focuses on the difference, such as it is, between the form of plural references in the Similar and Dissimilar conditions. The question can be phrased as follows: Do people economise on the content of references, opting for a logical conjunction by omitting TYPE or using a superordinate? A secondary question is whether, supposing the answer to the first question were negative, people perform some form of syntactic optimisation or aggregation, that is, whether they tend to produce descriptions of the form *the AP [*N_1 *and* N_2*]*, or whether the most likely form is *[the AP* N_1*] and [the AP* N_2*]*.

5.2.1 The form of plural referring expressions

Some examples of disjunctive and non-disjunctive descriptions from the PL_1 sample are shown in (5.5) and (5.6).

(5.5) (Disjunctive references)

	Plural Similar		Plural Dissimilar	
	Disjunctive	**Non-disjunctive**	**Disjunctive**	**Non-disjunctive**
aggregated	20.2	15.5	2.4	3.7
non-aggregated	64.3	–	93.9	–
% **of total**	84.5	15.5	96.3	3.7

Table 5.1: % disjunctive and non-disjunctive plural descriptions

(a) forward-facing red desk and fan

(b) the small red sofa and the small red desk

(c) the desk with its back to me and the grey sofa with its back to me

(d) the large red fan in the middle towards the left and the large red desk in the middle towards the centre

(5.6) (Non-disjunctive references)

(a) the two middle objects.

(b) two bottom blue

(c) the two leftmost objects in the middle row

(d) the two smallest red objects

Table 5.1 displays the percentage of descriptions in the two plural conditions, categorised according to whether they were *disjunctive* – involving a coordinate NP with two head nouns – or non-disjunctive, involving a simple NP with no TYPE attribute or a superordinate. The table also indicates whether disjunctive descriptions were syntactically aggregated with wide-scope modification as in (5.3a), or whether they were non-aggregated (that is, partitioned) like (5.2). To keep categorisation as conservative as possible, aggregated descriptions were defined as those with at least one property modifying two coordinate NPs. Thus, descriptions such as (5.7) below were considered as aggregated. In this case, *small* modifies both disjuncts.

(5.7) a red couch facing diaganally [sic] to the left and a desk facing diagannaly to the left which is red and both are small

As the table indicates, disjunctive descriptions were a majority in either condition, and most of these were non-aggregated. The majority contained basic-level TYPES (78.3%). In the Plural Similar condition, only 11.1% had no TYPE attribute at all, and even fewer used a superordinate term such as *object*, *furniture*, *item*, or *picture* (7.8%). Thus, the most likely reason for the majority of disjunctive descriptions in this condition is that people's descriptions represented a partition of a set of referents induced by the basic-level category of the objects. This conclusion is strengthened by two further results. First, there was no significant difference between the two plural conditions in the frequency of basic-level versus superordinate TYPE values ($\chi^2 = 5.354$, $p > .07$). Moreover, the likelihood of a description being disjunctive or non-disjunctive also did not differ as a function of Similarity ($\chi^2 = 2.56$, $p > .1$). The latter indicates that despite the fact that objects were visually identical in the Similar condition, except for the TYPE attribute, partitioned descriptions were the norm because the basic-level is preferred for this property.

A χ^2 test on overall frequencies of aggregated versus non-aggregated disjunctives showed that the non-aggregated, narrow-scope descriptions were a significant majority ($\chi^2 = 83.63$, $p < .001$). However, there was also a significant effect of Similarity on the frequency with which people aggregated their descriptions: the greater frequency of aggregation in the Similar condition compared to the Dissimilar (and the corresponding difference in frequency of non-aggregated descriptions in either condition) turned out to be significant ($\chi^2 = 15.498$, $p < .001$), despite the partitioned forms being a majority in both conditions.

The repetition of properties in narrow-scope disjunctive descriptions is somewhat surprising in view of proposals in the GRE literature, such as those by Gardent (2002), on the desirability of reducing epistemic redundancy. This criticism was levelled at van Deemter's IA$_{bool}$, which is prone to including non-contrastive properties in descriptions, and can also include a property more than once, because it occurs in several disjuncts (cf. §2.7.5, p. 58). The data suggests that a certain kind of epistemic redundancy is not viewed as problematic by the authors in the corpus. Whether this would turn out to facilitate reference resolution from a listener's point of view is a different question. However, Arts (2004, Ch. 4) did find that listeners' identification latencies were reduced when a description of an object was *exhaustive*, that is, it mentioned all the properties of the object (including non-contrastive ones). To the extent that it is possible to extrapolate from the singular to the plural case, this would imply that a disjunctive description that partitions a set and describes each element of the partition at the same level of detail will incur less effort in the reference resolution process.

The above conclusions about the form of plurals are somewhat tentative, since the data only involved reference to two objects, whereas reference to larger sets could conceivably give rise to a very different picture. However, on this dataset, it is likely that among the algorithms for plural reference generation reviewed in §2.7.5, the set partitioning strategy of van Deemter and Krahmer (2006) would best approximate the data as far as logical form is concerned, because the predominance of disjunctive descriptions without aggregation suggests that plurals correspond to partitions whose elements are described separately. However, this algorithm performs exhaustive search for a partition. Thus, it is non-incremental, and also lacks heuristics for maximising the adequacy of the content selected.

The partitioning of sets of referents when their elements have disjoint values of TYPE fits well with Pechmann's *gestalts* hypothesis, which makes TYPE central to the referential process because of the primacy of perceptual categorisation and the requirements of the syntactic module. Perceptual categorisation is a basic prerequisite to mental representation; indeed, without this capability, object recognition and classification would be impossible, as would generalisations about instances of classes or concepts (e.g. Murphy, 2002). Further motivation for the centrality of TYPE comes from psycholinguistically-oriented computational accounts of incremental syntactic formulation, the stage which follows (and is driven by) conceptualisation in the production pipeline of Levelt (1989). In the classic model proposed by Kempen and Hoenkamp (1987), syntactic phrase construction is head-driven and bottom-up. This means that phrases are constructed by mapping bits of conceptual structure to lexical items, which then project structure. Noun phrases in this model are therefore only constructed when a noun is available to function as the head of the phrase. The centrality of TYPE to the conceptualisation process guarantees that this can happen at the earliest possible stage, since this attribute is likely to be available early on in the incremental conceptualisation process, and is typically mapped to a noun. These theoretical considerations, in conjunction with the data on partitioning, can be summarised via a Principle of Category-driven Reference:

> **Category-driven Reference**
> The basic unit of the mental representation of an entity in a perceptual domain is its basic-level category, which is the product of perceptual categorisation and object recognition. This is also the basic input to the syntactic process of NP construction, whereby the concept or property corresponding to the category or TYPE of an object is mapped to a lexical item which is a noun, from which further structure in the NP can be projected.

If this is a correct generalisation, then any algorithm for reference should begin by categorising the intended referents. In the Incremental Algorithm, Dale and Reiter (1995) proposed a function that would insert TYPE at the end of the content determination process, if it was not selected because it lacked contrastive value. The Principle as stated above suggests that this should be the *first* step, which should take place irrespective of whether TYPE is contrastive. Moreover, the partitioning data suggests that in the case of plurals, object categorisation also determines the form of a plural referring expression. However, Pechmann's Gestalt model also gives primacy to other attributes, which are intimately bound to the conceptual representation of an entity. These are the attributes that Belke and Meyer (2002) refer to as those with high codability. As the analysis in Chapter 3 showed, attribute preferences are clearly in evidence in the data. The next section addresses some consequences of the interaction between people's tendency to partition sets along the lines induced by the basic-level TYPE of the referents, and the codability of attributes.

5.2.2 Codability and the Gestalt principle

To generalise Pechmann's observations about Gestalt representation to pluralities, a good starting point is the work on group perception that originated within the Gestalt school of psychology, particularly Wertheimer (1938). Wertheimer's principles hold that the ease with which a set of objects is perceived as a group is a function of (a) the **proximity** of the objects in the group, and (b) their **similarity** (see for example, Rock, 1983, for a discussion of these principles and their empirical verification). Since the focus of the present study is on domains in which objects were defined using perceptual attributes such as COLOUR, it is the Principle of Similarity that plays a central role. This principle predicts that the perception and conceptualisation of a set as a whole is facilitated if the elements of the set are perceptually similar.

If people's referring expressions reflect the way they are conceptualised, as Pechmann's Gestalts Hypothesis would predict, then the Similarity Principle would also predict that plural references should maximise the similarity between referents. Allowing for the independent motivation for set partitioning, based on TYPE values, Similarity predicts that descriptions such as *the red table and the red chair* are more likely than *the red chair and the large table*, since the latter does not use the same properties to describe the two referents. Similarity, however, should also interact with codability. High codability attributes are bound to feature in descriptions of sets for independent reasons. Given that descriptions of sets maximise similarity, and codable attributes are more easily perceived and represented, two related hypotheses emerge, which I will group together, referring to them as the *Parallel Structure Constraint* on partitioned descriptions:

> **Parallel Structure Constraint on Partitioning** (PSC)
> Descriptions of sets will tend to maximise the similarity between their elements, using the same attributes even when these are redundant. In particular:
>
> H1 (Codability) Highly codable attributes will feature redundantly in all elements of a partitioned description more often than non-highly codable attributes. Thus, in a partitioned description of length two, an attribute such as COLOUR, which has been found to have high codability, will be likely to be used twice in the description.
>
> H2 (Similarity) The use of repeated attributes in partitioned descriptions is more likely in the Similar, compared to the Dissimilar condition, because referents in the former are perceptually identical, save for their basic-level TYPE.

The investigation of the two predictions of PSC focused exclusively on the disjunctive descriptions in the PL_1 dataset ($N = 150$) from both Similar and Dissimilar conditions. The descriptions were categorised as follows:

1. **Descriptions with parallel semantics**: Disjunctive descriptions where the two coordinate NPs contain exactly the same attributes. Examples of these are shown in (5.8). For instance, (5.8c) contains two disjuncts (coordinate NPs), each of which identifies an entity based on TYPE, Y-DIMENSION and X-DIMENSION. Note that parallelism in attribute usage does not necessarily imply the same values across partitions. The values of the attributes are the same just in case a description was elicited in the Plural Similar condition.

   (5.8) (a) blue couch and green ventilator
   (b) small red chair facing front and large green desk facing front
   (c) top row desk on left and bottom row couch on left

2. **Descriptions with non-parallel semantics**: Disjunctive descriptions in which the two coordinate NPs do not contain exactly the same attributes. Examples are shown in (5.9). For instance, (5.9c) consists of two coordinated descriptions, both of which contain COLOUR and TYPE, but one contains Y-DIMENSION (*top row*), and one X-DIMENSION (*right-most*).

   (5.9) (a) the top red desk and the large grey sofa
   (b) green fan back large and small red chair top row
   (c) the grey dresser in the top row and the right-most grey chair

To find evidence for the PSC in the corpus, the primary focus should be on those descriptions that contain redundant information, that is, attributes which are not necessary to distinguish either referent. If they are included in both elements of a partitioned description, this would strongly

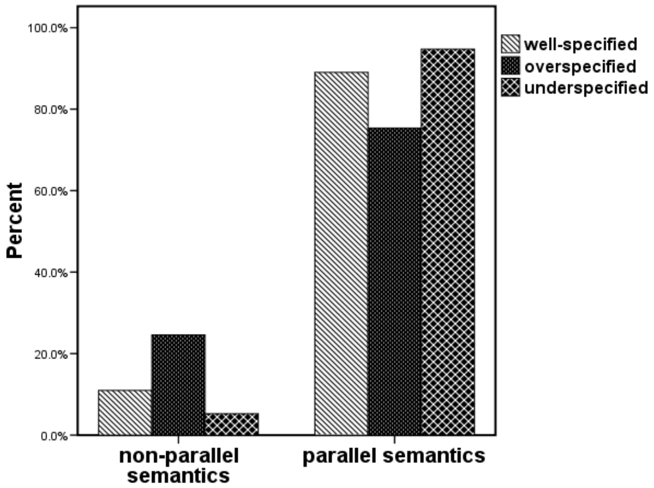


Figure 5.1: Parallelism in disjunctive descriptions

suggest that the PSC is on the right track. Therefore, the data analysis focused on evidence of parallel structure in overspecified descriptions, defined as per the definitions in §3.6.1 (p. 84), compared to underspecified and well-specified descriptions.

The analysis begins by looking at proportions of each of the above two types of description. These are displayed in Figure 5.1; the figures are given in Table 5.2.

	Non-Parallel	**Parallel**
overspecified	24.6	75.4
underspecified	5.3	94.7
well-specified	11	89

Table 5.2: Parallelism: % per description type

In all three description types, there is an overwhelming majority of descriptions that conform to the predictions of the PSC. This is confirmed by a χ^2 analysis, which showed that the majority was highly significant whether the description was well-specified ($\chi^2 = 92.467$, $p < .001$), overspecified ($\chi^2 = 42.217, p < .001$), or underspecified ($\chi^2 = 26, p < .001$). H1 above predicts that there should be a difference between Similar and Dissimilar conditions, with more evidence of parallel structure in the former. This, however, was not confirmed; the difference in proportions of description types across the two conditions was not significant ($\chi^2 < 1$, $p > .8$). Thus, no evidence in support of H2 was found. The data therefore suggests that the tendency to redundantly repeat attributes, avoiding aggregation, is independent of the Similarity of the elements of a set. This still leaves H1, which predicts that parallel semantic structure should be more clearly in evidence with highly codable attributes.

Once again, a test of H1 is strongest if performed on descriptions where the attributes in question are not required for a distinguishing description. I tested H1 by considering each of the 5 possible attributes in the TUNA corpus separately. Already, the analysis of Chapter 3 suggested an ordering among them, giving strong hints as to their status as highly codable attributes or not.

	Parallel (N)	Non-Parallel (N)	χ^2
SIZE	55.6 (5)	44.4 (4)	.111
COLOUR	77.4 (48)	22.6 (14)	18.645*
ORIENTATION	90 (9)	10 (1)	6.4**
X-DIMENSION	64.3 (27)	35.7 (15)	3.429
Y-DIMENSION	69.6 (39)	30.4 (17)	8.643*

Table 5.3: % Overspecified usage of attributes in parallel and non-parallel semantic structures. ($^*p < .005$, $^{**}p \leq .01$)

I compared proportions of parallel and non-parallel partitioned descriptions which contained a *redundant* use of each attribute. To take an example, H1 predicts that a highly codable attribute such as COLOUR is likely to be used in each element of a partition, even when it is not required. In the case of the three inherent visual attributes – COLOUR, SIZE and ORIENTATION – this was simply a matter of finding those descriptions which contained instances of these attributes when not required by the minimal description (MD) following the analysis in Chapter 3. For the two locative attributes, X-DIMENSION and Y-DIMENSION, which were not included in the MD calculation in the data elicitation experiment, I looked at proportions of overspecified descriptions containing LOCATION. Recall, from Chapter 3, that this class of descriptions consisted of those which included locatives, together with inherent visual attributes that were not part of MD.

Proportions of parallel and non-parallel descriptions for the five attributes are shown in Table 5.3. Because the sample used for analysis forms a relatively small proportion of the total corpus of plural descriptions, there were relatively few cases of overspecified usage of SIZE and ORIENTATION. The table includes absolute frequencies as well as percentages, and also shows the significant and non-significant χ^2 values, obtained by comparing frequencies of semantically parallel and non-parallel descriptions for each attribute. Although some caution is to be exercised in the case of SIZE and ORIENTATION, where the relevant data was sparse, the results are strongly supportive of H2, particularly because the trends parallel the results of §3.8 (p. 89). Overspecified use of COLOUR and vertical location (Y-DIMENSION) in plural descriptions was very likely to result in their inclusion in two disjuncts, corresponding to descriptions of each referent in the target set. The same goes for ORIENTATION. At $p = .08$, the results for X-DIMENSION failed to reach significance, again recalling the earlier result that this attribute tended to be used mostly in conjunction with Y-DIMENSION. In the case of SIZE, the difference in proportions failed to reach significance. This means that, even if this attribute was used when not required, it was not particularly likely to be included redundantly *twice* in a description. Thus, the data in the PL_1 sample contains as many descriptions like (5.10) as (5.11).

(5.10) (non-parallel)
the big red chair and the red desk with the drawer handles showing

(5.11) (parallel)
larger green fan, larger green chair

	Actual Values		**Predicted Values**	
	Singular	**Parallel**	**Linear**	**Non-linear**
COLOUR	.680	.835	.604	.61
SIZE	.290	.359	.283	.28
ORIENTATION	.280	.337	.269	.26
X-DIMENSION	.440	.706	.517	.52
Y-DIMENSION	.630	.899	.647	.65

Table 5.4: Actual and predicted probabilities of attribute usage

5.2.3 Further validation of the results

The trends reported above in support of H2 suggest that redundancy 'propagates' through parts of a description, so that an attribute which tends to form a central part of the representation of a referent is more likely to be used more than once in the same description. Given the finding that there was no significant difference between the Similar and Dissimilar conditions, this holds true even when the attribute repeated in two parts of a partitioned description has the same value.

If codability, or 'attribute preference', is indeed the determining factor in the phenomena under discussion, and if the observed trends have any generality, the likelihood with which an attribute is repeated in a plural partitioned description should be predictable from the likelihood with which it tends to be used overall, even in descriptions which are not plural. If this were found to be the case, it would lend stronger support to H1, suggesting that preferred attributes are repeated because this is a relatively easy option for a speaker or author. Such a result would also validate the trends reported earlier, which called for some caution in the case of some attributes due to data sparseness.

This was the rationale behind the test reported here. I used the descriptions in the TUNA corpus elicited in the *singular* condition ($N = 315$) to estimate the probability that a given attribute occurs in a description. These probability values were then used in a regression analysis to predict the likelihood of usage of an attribute in a partitioned description with parallel semantic structure. The relevant probabilities, obtained from the singular sub-corpus, are shown in the left panel of Table 5.4, together with the probability that the same attribute occur in a plural description with parallel semantics in PL_1.

A regression analysis was conducted to predict the probability of occurrence of an attribute in a parallel plural structure (denoted $p(\text{A}, \text{PPS})$) from its probability of usage in singular descriptions (denoted ($p(\text{A}, \text{SG})$). The analysis compared two regression models, which predicts the $p(\text{A}, \text{PPS})$ as a linear function of $p(\text{A}, \text{SG})$, and one which predicts the value as a non-linear, exponential function of $p(\text{A}, \text{SG})$. The resulting equations, obtained by fitting the parameters of the two models[4] to the data, are shown below. Equation (5.12) is the linear model, (5.13) the non-linear model.

$$p(\text{A}, \text{PPS}) = .042 + .673\, p(\text{A}, \text{SG}) \tag{5.12}$$

[4] The linear regression model has the form $p(\text{A}, \text{PPS}) = k + \beta\, p(\text{A}, \text{SG})$. The non-linear is of the form $p(\text{A}, \text{PPS}) = k\, p(\text{A}, \text{SG})^S$. In both, k is a constant intercept value.

$$p(\text{A}, \text{PPS}) = .713\, p(\text{A}, \text{SG})^{.912} \tag{5.13}$$

The predicted values for each attribute resulting from fitting the terms of the equations in each regression model are also shown in the right panel of Table 5.4. Both the linear and the non-linear model were highly reliable predictors of the likelihood of occurrence of an attribute in a parallel structure, based on its overall likelihood of occurrence in singular descriptions (linear: $\beta = .955$, $R^2 = .912$, $p = .01$; non-linear: $R^2 = .910$). This result strengthens the earlier conclusion that the overall preference of an attribute is indeed a good predictor of the likelihood of its usage in a description which is partitioned, where each element of the partition represents a subset of the set of target referents.

5.2.4 Interim summary

The foregoing analysis can be summarised as follows.

1. Human authors are likely to partition sets of referents along lines induced by the basic-level conceptual category to which the elements belong.

2. There is a strong tendency to describe elements of a partition in a parallel fashion, using the same attributes. This often results in considerable redundancy in descriptions.

3. This tendency is best viewed as a result of the relative ease with which specific attributes are processed in the incremental formulation of a description. High codability attributes are more likely to be repeated, and this is predictable from their overall probability of use.

Although the tendency to describe referents in parallel fashion was no greater when the referents had the same values on the relevant attributes, the results are nevertheless compatible with a weak version of the Similarity principle. What I have referred to as semantic parallelism is a way of conceptualising entities in the same way, using the same attributes as far as possible, though not necessarily with the same values. I call this a ‘weakened’ version of the Similarity principle because the overall similarity between elements of a partition is probably an emergent property of a description, whose cause seems to be the ease (or low cost) involved in the use of certain properties.

The notion of ‘low cost’ or ease of usage is reminiscent of the Krahmer and van der Sluis (2003) algorithm, reviewed in §2.7 (p. 47), which estimates the relative cost or effort involved in using an attribute in a description, as compared to a pointing gesture. In the present case, what the data suggests is that plural descriptions largely follow the trends observed in Chapter 3, with low-cost descriptive alternatives being included with greater likelihood. Some added complexity arises from the fact that once multiple referents are taken into account, the structure of the logical form is determined on the basis of how the entities are classified or categorised, with low-cost attribute being used several times in a partitioned description, giving rise to a conceptualisation of the referents that enhances their similarity. This result will turn out to be of some importance in the following two chapters. For now, the main task is to port the results to the GRE scenario. They give rise to the following desiderata:

1. An algorithm should observe constraints on *form*, producing (logical forms of) descriptions which mirror the partitioning strategies in the corpus data. Thus, partitions of a set of

	TYPE	COLOUR	ORIENTATION	SIZE	X-DIMENSION	Y-DIMENSION
e_1	desk	red	right	small	3	1
e_2	sofa	red	right	small	5	2
e_3	desk	red	back	small	1	1
e_4	desk	red	forward	large	2	3
e_5	desk	blue	right	large	2	4
e_6	sofa	red	back	large	4	1
e_7	sofa	red	forward	large	3	3

Table 5.5: A visual domain

referents should be induced by the basic-level TYPE of an object. Crucially, this result should be an emergent property of an incremental content determination strategy, rather than, say, a post-processing step which alters or normalises logical forms to make them meet certain criteria.

2. The algorithm should, whenever possible, observe the constraint on parallel structure, generating disjunctive references (where required) which conceptualise elements of a partition in a similar way, *if* this does not mean using dispreferred attributes.

5.3 Category-driven incremental generation by partitioning

The partitioning algorithm is presented here as a version of the IA (hereafter IA_{part}). I first discuss it using an informal example, and then formalise it with reference to the framework used in Chapter 2. Apart from the existence of the preference order, as in the original IA, the additional assumption will be made that every entity in the KB has one, and only one TYPE property. This is a simplification[5] which will be relinquished in later chapters.

5.3.1 An informal example

To motivate the development of IA_{part}, I will use the simple domain in Table 5.5. Suppose we require an algorithm to generate a reference to $\{e_5, e_6\}$. The corpus data suggests that authors are likely to partition this set by first selecting the category of the objects. In this case, this results in a partition of R into $\{\{e_5\}, \{e_6\}\}$ (*the sofa and the desk*).

A speaker would of course know that (a) this description is incomplete (it doesn't distinguish R); (b) each disjunct denotes one element of the partition of R, and is intended to refer to that element. Note that TYPE is a 'privileged' property, in that it is responsible for the initial partitioning. To make these things explicit, the description could be represented as a set of *fragments*, each corresponding to a disjunct. Each such fragment carries information about (a) which subset of R it is intended to refer to; (b) which property is the TYPE of that subset; (c) whether there are any other properties (roughly equivalent to modifiers of the TYPE). This representation – a triple consisting of these three elements – is shown below.

(5.14) $\langle\{e_5\}, \langle\text{TYPE} : \textit{desk}\rangle, \emptyset\rangle$

$\langle\{e_6\}, \langle\text{TYPE} : \textit{sofa}\rangle, \emptyset\rangle$

[5] This simplification is actually common in most of the GRE literature that deals with content determination. For example Dale and Reiter (1995) assume that there is only one TYPE attribute, to be mapped to the head noun of an NP, although their *findBestValue* function searches through a taxonomy of values to find the one closest to the basic level which is true of a referent and removes some distractors.

The representation says that there is a fragment which describes e_5, an entity which is categorised as a *desk*. The fragment has no other properties (yet). Similarly, mutatis mutandis, for e_6. Pending a precise definition of these fragments, I will stick to the convention of specifying the elements of the triples in this order (intended referents, type, and other properties).

At the stage represented by (5.14), an algorithm has the option of searching through disjunctive combinations of properties à la IA$_{bool}$, adding them to (5.14). This, however, would obscure the role of preferences for the attributes, as represented in the preference order. Consider instead a procedure that, having partitioned R by TYPE, made use of the same kind of 'divide-and-conquer' strategy, traversing a list of properties like the IA, and adding them to an element of the partition if relevant. This would impose two further requirements on the 'standard' picture of GRE. It would be necessary to structure a description as more than a set of properties, keeping track of which part of a partitioned description was intended to refer to which subset of R. Second, the algorithm has to keep track of which distractors have been removed for each element of the partition.

If these requirements are met, then the algorithm need only search through the preference order in the usual way. Suppose the first item in the attribute list is COLOUR. The algorithm would presumably find $\langle$COLOUR : *blue*$\rangle$ and add it to the disjunct denoting e_5, because this property removes some distractors for this referent, namely $\{e_1, e_3, e_4\}$. At this point, the description has the form in (5.15).

(5.15)
$$\Big\langle \{e_5\}, \langle \text{TYPE} : \textit{desk} \rangle, \big\{ \langle \text{COLOUR} : \textit{blue} \rangle \big\} \Big\rangle$$
$$\Big\langle \{e_6\}, \langle \text{TYPE} : \textit{sofa} \rangle, \emptyset \Big\rangle$$

According to the data analysis in the preceding sections, having added COLOUR to one part of the description, a human author is likely to add it to the other half, because this is a highly preferred property, and enhances the similarity between elements of the set. However, the property $\langle$COLOUR : *red*$\rangle$ has no contrastive value for e_6: all the distractors for this entity that remain after adding $\langle$TYPE : *sofa*$\rangle$ are red. What is needed at this point is a heuristic whereby an attribute is added if (a) it has been included in some other part of the description; and (b) it has high codability, or, correspondingly, relatively 'low cost'. There is, however, a potential pitfall. In the IA, the order in which properties are considered is determined by the preference order, but the order in which values of an attribute are considered is non-deterministic. In this case, for example, if the first value of COLOUR to be considered were *red*, it would be found to have no utility at all. Later, considering *blue*, the algorithm would include it because it is contrastive for e_5, but would have missed the chance of including *red* in the fragment corresponding to e_6, unless it performed some backtracking. The latter option trades off on incrementality, and is an undesirable feature for this reason, as well as the computational overhead it potentially incurs.

It seems reasonable to assume that, apart from their perceptual salience and centrality to mental representation, the perceptual contrastiveness of properties in a visual domain is also a determining factor in increasing the likelihood of their selection. This is the quality that van der Sluis (2005) has referred to as *inherent salience*. The inherent salience of an object depends on how many other objects in the domain have the same visual attributes. In Table 5.5, *blue* is highly contrastive with respect to e_5, since it is the only blue object. One way of getting around the problem is therefore to augment the attribute-driven ordering of properties in the IA with an ordering of the *values* of an attribute, by their discriminatory power, effectively combining the core features of the Incremental and Greedy algorithms. If this were achieved, then the algorithm in the current example would consider $\langle$COLOUR : *blue*$\rangle$ before $\langle$COLOUR : *red*$\rangle$, giving it a fighting chance of adding the latter property, assuming that an adequate heuristic were available to determine whether it has enough codability to warrant this. The description would now look like (5.16).

(5.16) $\Big\langle\{e_5\}, \langle \text{TYPE} : \textit{desk}\rangle, \big\{\langle \text{COLOUR} : \textit{blue}\rangle\big\}\Big\rangle$
$\Big\langle\{e_6\}, \langle \text{TYPE} : \textit{sofa}\rangle, \big\{\langle \text{COLOUR} : \textit{red}\rangle\big\}\Big\rangle$

There is a complication which has so far not been considered. The data on which the preceding analysis is based consists entirely of references to two objects. What of references to three or more referents? In case the referents belonged to different categories (had different TYPE values), the strategy uncovered in the data would generalise easily: a fragment of the description would correspond to each referent. More problematic is the case where, say, two referents have the same TYPE, but then have different values of some other attributes. For example, suppose that $R = \{e_2, e_3, e_4\}$. Partitioning by TYPE results in two fragments, corresponding to the sofa (e_2) and the two desks ($\{e_3, e_4\}$). Adding COLOUR, to yield the equivalent of *the red desks and the red sofa* still does not do the trick, because e_1 is a distractor which is a red desk. Suppose ORIENTATION is considered next. This will distinguish the two desks from e_1. However, they have different values of this attribute. Since the hypothetical algorithm proceeds incrementally, considering each value of ORIENTATION in turn, one possibility would be to allow *any* property to induce a partition on R. Thus, the subset $\{e_3, e_4\}$ would be broken up even further. On encountering $\langle$ORIENTATION : *backward*$\rangle$, the description would consist of three fragments, as follows:

(5.17) $\Big\langle\{e_2\}, \langle \text{TYPE} : \textit{sofa}\rangle, \big\{\langle \text{COLOUR} : \textit{red}\rangle\big\}\Big\rangle$
$\Big\langle\{e_3\}, \langle \text{TYPE} : \textit{desk}\rangle, \big\{\langle \text{COLOUR} : \textit{red}\rangle, \langle \text{ORIENTATION} : \textit{backward}\rangle\big\}\Big\rangle$
$\Big\langle\{e_4\}, \langle \text{TYPE} : \textit{desk}\rangle, \big\{\langle \text{COLOUR} : \textit{red}\rangle\big\}\Big\rangle$

The next value of ORIENTATION, *right*, would then be added to the fragment for e_4. This kind of strategy is adopted here. However, having a description consisting of three fragments, two of which have the same TYPE, runs counter to the evidence for category-driven reference. Clearly, there will be cases where limits on syntactic complexity of descriptions would interact with the category-driven constraint, and may well result in coordinate NPs with identical head nouns because the alternative would be far too complex. However, to the extent that the results of the previous analysis are generalisable, they would suggest that aggregation would be performed, to respect the constraint that partitions occur primarily by category. In the current example, it is possible to aggregate the two fragments corresponding to e_3 and e_4 , yielding the equivalent of *the red desks, one facing backward and the other facing forward.* Furthermore, this can be carried out opportunistically. If the algorithm kept track of which fragments referred to which subset of R, aggregation could be triggered as soon as a fragment is complete, that is, as soon as it distinguished the referents to which it corresponded.

This problem is guaranteed not to occur in the kinds of domains under discussion, where the set of intended referents never exceeded two. For the present, I will put it aside and describe the content determination procedure, returning to it in §5.5.

Before turning to a formalisation of the algorithm, some properties of the data structures that I have referred to as fragments, as well as the procedure that gave rise to them, are worth highlighting:

1. Any two fragments are intended to refer to disjoint subsets of R.

2. Fragments represent logical conjunctions of properties. For example, the fragment corresponding to e_3 in (5.17) is equivalent to the description *the red desk facing backward*.

3. Because each fragment is intended to refer to some subset of R, the procedure just sketched involves breaking down the intention to refer to a plurality (R) into a number of 'sub-intentions' to refer to its subsets. In the foregoing example, these 'sub-intentions' were formulated on the fly, based on the properties considered (for instance, consideration of ORIENTATION resulted in the partition represented by (5.17).

To mimic the foregoing examples, the algorithm described below starts out by finding values of TYPE for the referents. This is also the stage at which fragments are initially constructed. In the case of $R = \{e_5, e_6\}$, for example, this initial step finds two values of TYPE to yield (5.14). Subsequently, the algorithm traverses the ordered list of properties in the by-now familiar way. Any property that is true of *at least one* element of R is a candidate for inclusion. (This distinguishes it from IA$_{bool}$ and IA$_{plur}$ (van Deemter, 2000), which both require that a property or combination thereof be true of the entire set.) Let R' be the referents of which some property $\langle$A : *v*$\rangle$ is true. The property will be included if either one of the following conditions hold:

1. $\langle$A : *v*$\rangle$ is contrastive with respect to R', that is, it excludes some distractors for the elements of this set.

2. The attribute A has already been used in the description, and it is sufficiently highly preferred to warrant the (redundant) inclusion in the description of $\langle$A : *v*$\rangle$. In the preceding example, such a case arose with $\langle$COLOUR : *red*$\rangle$, which denotes e_6, but does not exclude any distractors for it. As noted in the earlier discussion, to be able to include this property when it is non-contrastive, the property $\langle$COLOUR : *blue*$\rangle$, which *is* contrastive for e_5, will have to have been included before *red* is considered. Therefore, values of attributes are considered using the greedy heuristic.

If a property satisfies either of the above conditions, then the description is updated. To do this, the main thing to consider is which referents the property is true of (i.e. the set R'). The update of a description with a new property can have either of two consequences. Suppose there is a fragment whose intended referents are subsumed by R'. Then the property is simply added to that fragment. This is what happened, for example, when $\langle$COLOUR : *red*$\rangle$ was included in (5.17). Another possibility is that the new property induces a further partitioning on a fragment. This is exactly what happened when $\langle$ORIENTATION : *backward*$\rangle$ was considered earlier. This was only true of e_3, so that the fragment which was intended to refer to $\{e_3, e_4\}$ was further partitioned to yield (5.17).

5.3.2 Definitions and assumptions

This section makes the intuitive outline of the algorithm more precise. I begin by defining the data structures used by IA$_{part}$, called **Description Fragment**s (DF).

Definition 5. Description Fragment
A Description Fragment (DF) is a triple $\langle R_{\text{DF}}, T_{\text{DF}}, M_{\text{DF}} \rangle$ where:

- $T_{\text{DF}} \in P_R = \langle A, V \rangle : A = \text{TYPE};$
- $R_{\text{DF}} \subseteq [\![\, T \,]\!];$
- $M_{\text{DF}} \subseteq P_R = \{ p \mid [\![\, p \,]\!] \cap [\![\, R_{\text{DF}} \,]\!] \neq \emptyset \}$
- $[\![\text{ DF }]\!] = \bigcap_{p \in M} [\![\, p \,]\!] \cap [\![\, T_{\text{DF}} \,]\!]$

In other words, DFs are conjunctions of a TYPE and possibly some other properties. I now define a **Partitioned Description**. Apart from its definition as a set of DFs, it is also required that no two DFs be intended to refer to the same elements of R.

Definition 6. Partitioned Description
A Partitioned Description (D_{part}) is a set of description fragments where:

- $\forall \text{DF}, \text{DF}' \in D_{part} : R_{\text{DF}} \cap R'_{\text{DF}} = \emptyset$
- $[\![\, D_{part} \,]\!] = \bigcup_{\text{DF} \in D_{part}} [\![\text{ DF }]\!]$

By this definition, DFs represent partitions, so that no two DFs have the same set of intended referents. The description is equivalent to a formula in Disjunctive Normal Form (DNF) and, extensionally, is the union of the extensions of the fragments making it up.

Next, I define some of the basic ingredients for a full description of the algorithm, along the lines of my earlier formalisation in §2.4 (p. 31). Recall that in order for a property to be considered for inclusion, it suffices that it be true of some subset of R. This means that P_R, the set of *relevant properties* through which IA$_{part}$ searches, is defined as follows, where $\mathbb{P}$ is the set of properties in the KB (compare to Definition 4, p. 59).

$$P_R = \{ p \mid p \in \mathbb{P} \wedge R \cap [\![\, p \,]\!] \neq \emptyset \} \tag{5.18}$$

As per the preceding discussion, the search procedure adopted in IA$_{part}$ is based on an ordering relation among properties, whereby preferred attributes are considered first, and preferred values of a given attribute are prioritised using the Greedy heuristic (Dale, 1989). Let $disc(p)$ abbreviate 'the discriminatory power of property p'. In Chapter 2, this was simply defined as the number of distractors which a property excludes, an adequate definition given that P_R was defined as all those properties in the KB which have R in their extension. In view of the revised definition of P_R, $disc(p)$ is now defined as a function of the number of referents that a property has in its extension, and the number of distractors it excludes. The value, a real number in $(0, 1)$ (where 1 indicates the maximal discriminatory power) is calculated as in (5.19), where $C = U - R$.

$$disc(p) = \frac{|[\![\, p \,]\!] \cap R| \; + |[\![\, p \,]\!] - C|}{|[\![\, p \,]\!]|} \tag{5.19}$$

Turning next to the ordering relation that structures the search space of the algorithm, let PO be the predefined ordered list of attributes, and $index(\textsc{a}, \textsc{po})$ be the position of attribute A in PO. Formally, the ordering relation $>>_{p_{part}}$ between properties is defined as follows:

$$\langle \textsc{a} : v \rangle >>_{p_{part}} \langle \textsc{a'} : v \rangle \leftrightarrow \begin{cases} index(\textsc{a}, \textsc{po}) < index(\textsc{a'}, \textsc{po}) \text{ if } \textsc{a} \neq \textsc{a'} \\ disc(\langle \textsc{a} : v \rangle) > disc(\langle \textsc{a'} : v \rangle) \text{ otherwise} \end{cases} \tag{5.20}$$

Therefore, a property precedes another property if the attribute of which it is a value is found earlier in the preference order. In case two properties represent values of the same attribute, they are ordered with respect to their discriminatory power. For the sake of consistency with earlier formalisations of GRE algorithms, I also define the return value of the $dequeue(Q)$ function. This is the function that returns the next property in the priority queue held by a GRE algorithm (cf. §2.4.1, p. 32). Let $values(\textsc{a})$ denote the values of an attribute A. Then:

$$\begin{aligned} dequeue(Q) =_{def} \langle \textsc{a} : v \rangle : \textsc{a} = &\operatorname*{arg\,min}_{\langle \textsc{a'}:v' \rangle \in P_R} index(\textsc{a'}, \textsc{po}) \wedge \\ \textsc{v'} = &\operatorname*{arg\,max}_{v' \in values(\textsc{a'})} disc(\langle \textsc{a'} : v' \rangle) \end{aligned} \tag{5.21}$$

Two further ingredients for the algorithm are required. The first is a revised notion of a 'distractor set'. I will assume that the algorithm has at its disposal an associative array C, holding a set of distractors *for each element of* R. Thus, $C[r]$ for some $r \in R$ is the set of distractors of r given the current state of the content determination procedure. In what follows, the notion of contrastiveness of a property is abbreviated by the boolean function $contrastive(p)$, which is defined as follows (cf. Krahmer and Theune, 2002, for a similar use of this function).

$$contrastive(p) \leftrightarrow \exists r \in R : \; C[r] - [\![\, p \,]\!] \neq \emptyset \tag{5.22}$$

The final ingredient is required to ensure that the algorithm maximises the similarity between partitions or DFs, by propagating an attribute across fragments, if that attribute has high codability (i.e. is sufficiently highly preferred). This will occur in case a property is true of some referents but has no contrastive value for them. Such a property is added if it will enhance the similarity (the Gestalt status) of the intended referents, and has sufficiently high codability (low cost) to be included redundantly. I use the regression equation in (5.13) to operationalise the notion of codability; this was found to be highly predictive of the likelihood with which an attribute is used in a parallel structure, and is based on usage probabilities derived from singular data. It forms the basis for the definition of another Boolean function, $useful(\langle \textsc{a} : v \rangle, D_{part})$, which is an abbreviation for 'the property $\langle \textsc{a} : v \rangle$ is useful with respect to D_{part}'. This is defined below, where $att(D_{part})$ is the set of attributes in the description:

$$useful(\langle \textsc{a} : v \rangle, D_{part}) \leftrightarrow \textsc{a} \in att(D_{part}) \wedge \left(.713 \, p(\textsc{a}, \textsc{sg})^{.912} > 0.5 \right) \tag{5.23}$$

In other words, an attribute-value pair is useful with respect to the description if (a) the attribute is already represented in at least one DF, and (b) its likelihood of being propagated across elements of a partitioned description is greater than chance.

5.3.3 Formalisation of the algorithm

Pseudocode for IA$_{part}$ is displayed in Algorithm 3. For convenience, it is divided into three main procedures, all of which utilise some 'global' variables: the description D_{part}, initially set to empty [3.2], and the priority queue, which is intialised to contain all properties in P_R [3.1].

The procedure *makeReferringExpression* [3.3–3.16] is the main function, which searches through the set of relevant properties. Prior to search, it calls the procedure *makeTypes* [3.17–3.24] , which initialises the description by finding TYPE values for each referent, removing them from the queue to avoid them being considered again later [3.20–3.21]. It is crucial that the main procedure *makeReferringExpression* initialise the description by calling *makeTypes*, because this procedure searches for TYPE values [3.18], and updates the description with these properties [3.20]. This results in an initial partitioning according to the conceptual category of the referents. Note, moreover, that following the initial call to *maketypes*, the intended referents of the DFs are determined by the TYPE of the referents in R.

Both procedures make use of the procedure *updateDescription* to add properties to the description D_{part} [3.25–3.47]. This procedure takes as arguments the subset of R (R') of which a property is true, and the property itself. It is this procedure that maintains the partitioned structure of a description, and it is most useful to turn immediately to a discussion of how it works. This procedure consists of a main for loop [3.26–3.39], and a final condition [3.40–3.46]. The basic idea is the following: *Given a property and the set of referents it is true of, try to find one or more fragments in the description in which that property can be included, until the entire set R' is accounted for*, that is, there is a fragment containing every element of R', which also includes the new property.

In the main for loop, the algorithm iterates through D_{part}. At each point, it checks whether the intended referents R_{DF} of a fragment include at least some elements of R'. Recall, from the informal discussion, that a new property can either be added to a fragment, or cause a further partitioning of the fragment. These two cases are specified as follows:

1. R' includes all the intended referents of a DF [3.29]. In this case, the procedure adds the new property to the set M_{DF} in this DF [3.30]. R' is then updated by removing the referents thus accounted for [3.31]. For example, suppose the algorithm were called with $R = \{e_5\}$ from Table 5.5. *makeTypes* will return a single DF consisting of e_5 and $\langle$TYPE : *desk*$\rangle$. On encountering $\langle$COLOUR : *blue*$\rangle$, the function *makeReferringExpression* calls the update procedure with this property and $R' = \{e_5\}$. This is simply added to the sole DF in the description, as $R' = R_{\text{DF}}$, which satisfies the condition at [3.29];

2. Some, but not all, of the referents included by the property are also referred to by the DF, that is $R' \cap R_{\text{DF}} \neq \emptyset$ [3.32]. In this case, the DF is partitioned into two:

    (a) A new DF is created whose intended referents are $R' \cap R_{\text{DF}}$ [3.33]. This DF inherits the TYPE of the original DF, and also all of its modifiers, together with the new property [3.35]. The new DF is added to the description [3.36]

Algorithm 3 IA_{part}

```
Require: $R$, $P_R$   ▷ input: referents and their relevant properties
$Q \leftarrow$ a Priority Queue, containing all properties in $P_R$   ▷ a global variable
$D_{part} \leftarrow \emptyset$   ▷ the description, a global variable

procedure makeReferringExpression   ▷ main procedure
    $makeTypes()$   ▷ add TYPEs; description is partitioned
    while $Q \neq \emptyset$ do
        if $[\![ D_{part} ]\!] = R$ then   ▷ terminate if $R$ is distinguished
            return $D_{part}$
        end if
        $\langle \text{A} : v \rangle \leftarrow dequeue(Q)$   ▷ next property
        if $contrastive(\langle \text{A} : v \rangle) \vee useful(\langle \text{A} : v \rangle)$ then   ▷ property must be contrastive or warrant redundant propagation
            $R' \leftarrow R \cap [\![ \langle \text{A} : v \rangle ]\!]$   ▷ initialise set of referents included in current property
            $updateDescription(R', \langle \text{A} : v \rangle)$   ▷ update the description
        end if
    end while
    return $D_{part}$   ▷ $R$ has not been distinguished
end procedure

procedure makeTypes   ▷ adds TYPE to $D_{part}$, partitions if necessary
    for $\langle \text{A} : v \rangle \in P_R$ do   ▷ any property in $P_R$ is true of some $r \in R$, by definition
        if $\text{A} = \text{TYPE}$ then
            $D_{part} \leftarrow updateDescription([\![ \langle \text{A} : v \rangle ]\!] \cap R, \langle \text{A} : v \rangle)$   ▷ create new fragment
            $Q \leftarrow Q - \{\langle \text{A} : v \rangle\}$   ▷ remove the TYPE from $Q$
        end if
    end for
end procedure

procedure updateDescription($R'$,$\langle \text{A} : v \rangle$)
    for $\langle R_{\text{DF}}, T, M \rangle \in D_{part}$ do
        if $R' = \emptyset$ then   ▷ terminate as soon as $R'$ accounted for
            return
        else if $R_{\text{DF}} \subseteq R'$ then   ▷ property is true of all elements in this DF
            $M \leftarrow M \cup \{\langle \text{A} : v \rangle\}$   ▷ no partitioning required
            $R' \leftarrow R' - R_{\text{DF}}$   ▷ remove referents accounted for from $R'$
        else if $R_{\text{DF}} \cap R' \neq \emptyset$ then   ▷ property is true of some elements of the DF
            $R_{new} \leftarrow R_{\text{DF}} \cap R'$   ▷ $R_{new}$ contains referents of which the property is true
            $R_{\text{DF}} \leftarrow R_{\text{DF}} - R_{new}$   ▷ update $R_{\text{DF}}$ to ensure that DFs represent partitions
            $DF_{new} \leftarrow \langle R_{new}, T, M \cup \{\langle \text{A} : v \rangle\} \rangle$   ▷ new DF has all the old properties, plus the new one
            $D_{part} \leftarrow D_{part} \cup \{DF_{new}\}$   ▷ update the description with the new DF
            $R' \leftarrow R' - R_{new}$   ▷ remove referents accounted for from $R'$
        end if
    end for
    if $R' \neq \emptyset$ then
        if $\text{A} = \text{TYPE}$ then   ▷ special case: this is a TYPE
            $D_{part} \leftarrow D_{part} \cup \{R', \langle \text{A} : v \rangle, \emptyset\}$
        else   ▷ not a TYPE (KB is incomplete)
            $D_{part} \leftarrow D_{part} \cup \{\langle R', \bot, \{\langle \text{A} : v \rangle\} \rangle\}$
        end if
    end if
end procedure
```

(b) The original DF is updated to reflect the split, so that R_{DF} no longer contains the elements in the new fragment [3.34]. This ensures that fragments always represent partitions.

The way descriptions are updated at this stage therefore ensures that each fragment in the description satisfies Definition 6, since no two DFs will describe the same intended referents. However, by case (2) above, it is possible for a property to result in a partition that is not category-driven. This case corresponds to the example discussed earlier, where two referents with the same TYPE, for example $\{e_3, e_4\}$ in Table 5.5, have different values of another attribute.

In case D_{part} is empty, or there are insufficient DFs to account for R', this set will not be empty at the end of the loop. The final condition at line [3.40] deals with this case, which once again gives rise to two possibilities, this time depending on the attribute under consideration:

1. If the property is a value of TYPE, the algorithm will construct a new DF consisting of what remains of R', the property itself (which corresponds to T_{DF} in the DF, and an empty set of modifiers [3.42]. This is essentially what happens at the beginning of the main procedure, when the description is empty, and *makeTypes* calls *updateDescription* with new values of TYPE.

2. If the property is not a TYPE, then the algorithm constructs a new DF at [3.44], consisting of R', a null ($\perp$) value for TYPE, and the set M_{DF} consisting of the property under consideration. This only happens if the assumption made earlier, namely that all entities have a TYPE, is violated by the Knowledge Base. Note, however, that relinquishing this assumption will not change the character of the algorithm, since descriptions are still partitioned.

The main procedure *makeReferringExpression* consists of a $while$ loop [3.5], whereby properties in the queue are dequeued [3.9]. A property is added to the description if either (a) it is contrastive in the sense defined in (5.22) or (b) it is useful in the sense of (5.23). Satisfaction of either condition results in a call to *updateDescription* [3.12]. The loop terminates as soon as the description is distinguishing [3.7].

Because the utility (or codability) of a property is taken into account, IA$_{part}$ can sometimes return a description which is more overspecified than it would be were the standard IA run over the same domain. Another source of extra overspecification stems from the fact that the function $contrastive(p)$ as defined in (5.22) evaluates to `true` if there is at least one referent for which that property has contrastive value. It is conceivable that the property be true of a number of referents, but have contrastive value only for some of these. This is effectively another way of enhancing the parallel semantic structure of the description. Apart from this, ordering the properties in P_R by attribute *and* by discriminatory power guarantees that every value of an attribute that is dequeued [3.9] is only redundantly included if another value of the same attribute has already been included.

5.3.4 Complexity of IA$_{part}$

An estimate of the complexity of the procedure described above needs to take two factors into account:

1. The maximal number of iterations of the main $while$ loop of *makeReferringExpression*. This is clearly bounded by $n_p = |P_R|$.

2. The maximal number of iterations of the for loop in *updateDescription*, whereby the algorithm adds a new property to D_{part}. This is bounded by $n_r = |R|$. To see this, consider the two cases where a new DF is constructed. First, DF construction can triggered by a TYPE property. Since, by assumption, every referent has at most one TYPE value, this will maximally return $|R|$ DFs. Second, a DF can be constructed when an existing fragment is partitioned. Since there are as many DFs as there are disjoint TYPEs for the referents, this can only happen if there are referents which share a TYPE value. Therefore, further partitioning cannot result in more than $|R|$ DFs.

This gives the algorithm a worst-case runtime complexity $O(n_p n_r)$. Arguably, another factor needs to be taken into account. Because the queue is assumed to be dynamic, it will have to keep track of when distractors have been removed, recalculating the discriminatory power of properties every time this happens. (This is not explicitly shown in Algorithm 3.) Here, we can mirror an argument by Dale and Reiter (1995): Suppose there are n_d unique properties in the description returned by the algorithm. Only unique properties need to be considered, as the number of times a property is included in the description does not affect the number of times the queue is to be updated. This means that in order to return the most discriminatory value of an attribute, the algorithm has to test n_p properties at most n_d times. Overall, this gives IA$_{part}$ complexity $O(n_p^2 n_r n_d)$.

The polynomial-time complexity of the algorithm contrasts with some earlier proposals for plural GRE. In particular, IA$_{bool}$ (van Deemter, 2002) loses the polynomial complexity of the original IA because it enqueues disjunctions. The constraint-based algorithm proposed by Gardent (2002) constitutes a return to Dale's (1989) Full Brevity heuristic, a known intractable problem. The partitioning strategy adopted here is also more efficient than that proposed by van Deemter and Krahmer (2006). This algorithm searches through successive partitions of R until one is found whose elements can be described non-disjunctively. In the worst case, this algorithm will need to search through all partitions of a set, making it exponential in $|R|$. While the present strategy has much in common with this one, in that it too seeks to describe subsets of R non-disjunctively, combining several fragments of a description into a single disjunction, it performs partitioning opportunistically, based on the properties in the KB.

The calculation of complexity of IA$_{part}$ has not taken *full* Boolean completeness into account, in that negation has not been explicitly treated. However, van Deemter (2002) showed that negation can be handled relatively easily. It increases the number of properties (the size of P_R in the current terminology) by a factor of 2, because the negation of every literal is also added. If this is assumed to be an offline or pre-processing task, then the theoretical complexity of the algorithm remains unaltered. Beyond the purely formal details, however, adding negation raises a large number of empirical questions which have yet to be investigated, and which go beyond the scope of the present work. I will however return briefly to one possible use of negation in the concluding section.

5.4 Evaluating the partitioning algorithm

The evaluation of IA_{part} took the form of a comparison of its performance against that of IA_{bool}, using the subset PL_2 of the corpus to compare the output of the two algorithms against human-authored descriptions. To maximise the similarity of the two algorithms, IA_{part} was implemented as an extension to the GRE-API introduced in Chapter 4, maintaining the assumptions made so far about the algorithms.

As in the previous chapter, the descriptions in PL_2 were divided into a −LOC and a +LOC dataset. However, the division took a slightly different form. In §4.5 (p. 120), the analysis showed that inconsistency among authors in their use of locative attributes resulted in significant variability in the performance of the IA on the dataset containing locatives (see especially Figure 4.4). Therefore, in the present study, rather than divide the data by the condition in which authors wrote their descriptions (i.e. whether they were allowed to use locative expressions), I divided them according to whether the descriptions actually contained a locative or not. As a result, the datasets are not balanced by items or subjects, in the sense that it is no longer guaranteed that the number of authors in a given dataset, and the number of domains for each author, are approximately equal. Therefore, the analysis will report two-tailed t-tests averaging over all the descriptions within a dataset, rather than authors and/or domains. In this study, the +LOC dataset consisted of 148 descriptions, while −LOC consisted of 257.

Since the purpose of this evaluation was to compare two alternative strategies for dealing with disjunction and plurality, rather than comparing the output of several versions of the IA, attention was restricted to those preference orders found to perform best on −LOC and +LOC data, once the variability of authors is accounted for as in §4.5. The orders selected are shown below; note that the order used for +LOC is one of the best-performing orders *when evaluated only on locative descriptions*.

(5.24) (+LOC)
Y-DIMENSION >> COLOUR >> X-DIMENSION >> SIZE >> >> ORIENTATION
(−LOC)
COLOUR >> ORIENTATION >> SIZE

5.4.1 Evaluation functions

The main purpose of the evaluation was to assess whether the partitioning strategy would yield a better match to the human data both on the content of referring expressions, and on their form. In order to assess agreement on content, the Dice coefficient, as described in §4.3.2 (p. 110), was retained as an evaluation function. A more conservative measure, one based on Levenshtein (‘edit’) distance, was also used. Dice focuses exclusively on the extent to which two descriptions contain the same attributes the same number of times[6], and a score below 1 reflects whether an algorithm contained attributes that were not in the human-authored description, and also whether it failed to include attributes that were. This measure will not take into account the syntactic aspects of the logical form generated by an algorithm, compared to that produced by a human. For example, the two formulae in (5.25) will be given exactly the same score compared to a human description, because they contain exactly the same attributes.

[6] This is the version of Dice used in the previous chapter, with descriptions represented as multisets.

(5.25) (a) $\left[\langle \textsc{type} : \textit{chair}\rangle \wedge \langle \textsc{colour} : \textit{blue}\rangle\right] \vee \left[\langle \textsc{type} : \textit{fan}\rangle \wedge \langle \textsc{colour} : \textit{red}\rangle\right]$

(b) $\left[\langle \textsc{type} : \textit{chair}\rangle \vee \langle \textsc{type} : \textit{fan}\rangle\right] \wedge \left[\langle \textsc{colour} : \textit{blue}\rangle \wedge \langle \textsc{colour} : \textit{red}\rangle\right]$

A measure that also took into account the form of expressions would give different scores, depending on whether the gold standard to which they are compared contains exactly the same logical operators in exactly the same position. For this reason, I included Levenshtein ('edit') distance as an evaluation measure. The classic version of edit distance (Levenshtein, 1966) compares two strings, finding the minimum cost of transforming one into the other. Cost is defined in terms of additions, deletions and substitutions. Since the GRE-API represents formulae as trees, the calculation of edit distance for this experiment used the tree distance algorithm proposed by Shasha and Zhang (1990). Let i, d, and s be the predefined cost of performing an insertion, a deletion and a substitution respectively. Let t_1 and t_2 be two ordered trees, r_t the rightmost non-leaf node of any tree t, and $T(r)$ the tree rooted at node r. The Shasha and Zhang algorithm generalises the definition of distance $\delta(t_1, t_2)$ as shown below.

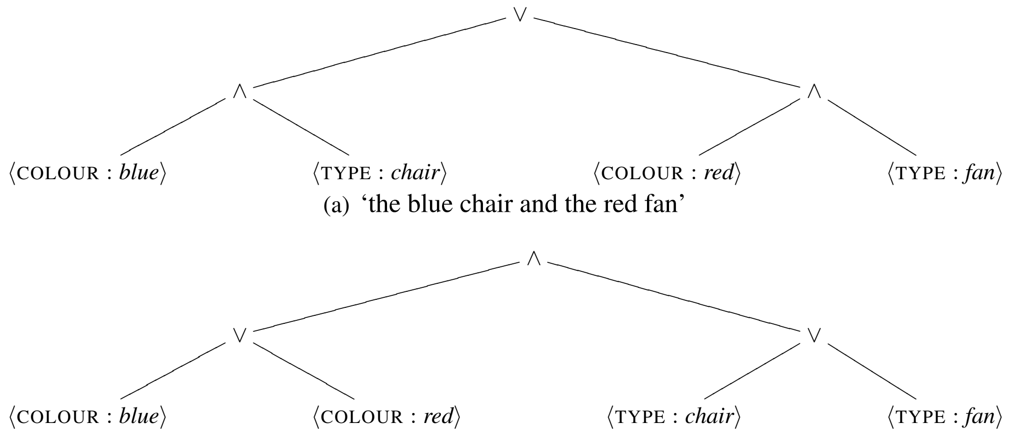
$$\begin{aligned}
\delta(\bot,\bot) =& 0\\
\delta(t_1,\bot) =& \delta(t_1 - r_{t_1},\bot) + d\\
\delta(\bot,t_2) =& \delta(\bot,t_2 - r_{t_2}) + i\\
\delta(t_1,t_2) =& \min\begin{cases}\delta(t_1 - r_{t_1},t_2) + d\\ \delta(t_1,t_2 - r_{t_2}) + i\\ \delta(T(r_{t_1}),T(r_{t_2})) + \delta(t_1 - T(r_{t_1}),t_2 - T(r_{t_2})) + s\end{cases}
\end{aligned} \tag{5.26}$$

The edit distance between two trees ranges between 0 and ∞, where a value of 0 indicates identity. For the purposes of this evaluation i and d were both set to 1, while s was set to 2. The distance between two formulae was computed by walking the trees in left-to-right pre-order; thus, a node's children were always deleted before the node itself. Because this measure is applicable to ordered trees[7], sibling attribute-value nodes were always ordered in an arbitrary but fixed order within the formula (e.g. TYPE always preceded COLOUR, and an order was defined within values of each attribute).

To take an example of how tree edit distance is calculated, assume that (5.25a) is being compared to (5.25b). The tree representation of these formulae is shown in Figure 5.2.

To transform the tree in (5.2(b)) into (5.2(a)), the topmost non-terminal nodes ($\vee$ and two occurrences of $\wedge$) have to be substituted, resulting in a cost of ($3 \times 2 =$) 6. Moving from left to right through the non-terminals, the first TYPE node in (5.2(b)) is left intact, but the second needs to be substituted for a COLOUR property. The substitution incurs another cost of 2. The second conjunct in the tree requires the deletion of the leftmost COLOUR node, and the insertion of a TYPE node, incurring another cost of 2, and giving a total cost of 10. This is in stark contrast to the score given by the Dice coefficient for the same comparison, which will give a cost of 1 (identity) because the trees contain exactly the same properties.

[7] Tree edit distance for unordered trees is a known NP-Hard problem.



(a) ‘the blue chair and the red fan’

(b) ‘the blue and red chair and fan’

Figure 5.2: Tree representations of two disjunctive formulae

As in Chapter 4, the evaluation functions only considered attributes (and the operators $\vee$ and $\wedge$ in the case of edit distance), as opposed to attribute-value pairs. This was to avoid penalising the algorithms in those cases where *other* values were used by humans, which were not present in the domains to which the algorithms were exposed.

5.4.2 Results and discussion

Dice coefficient and Levenshtein scores for the two algorithms are shown in Table 5.6. As in the earlier evaluation, the table displays mean and modal scores, as well as the perfect recall percentage (PRP), the proportion of 1 Dice scores and 0 Levenshtein scores. The mean scores per dataset on the two measures are also displayed in Figure 5.3.

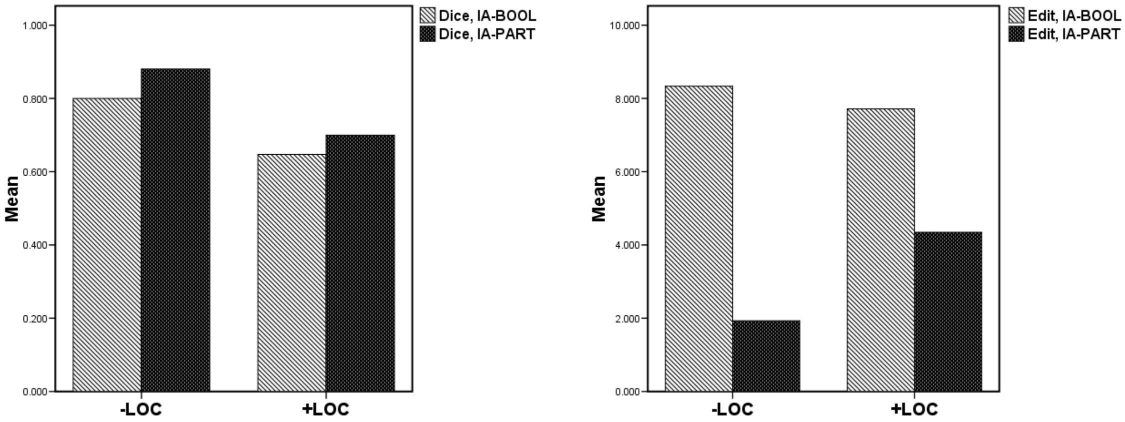
The overall trends in the data across the two datasets mirror those found in the previous chapter, with lower performance for both algorithms in the $+$LOC dataset. However, IA$_{part}$ performs better on both datasets, on both measures. The modal score of this algorithm on the $-$LOC dataset is 1 on Dice and 0 on Levenshtein distance, meaning that it matched human descriptions on both form and content perfectly most of the time. On $+$LOC, IA$_{part}$ obtained a PRP of 6.8 on Dice and edit, compared to 1.4 and .7 for IA$_{bool}$. Note that the PRP on both measures is always identical for IA$_{part}$, suggesting that when it agreed perfectly on content with an author, as measured by Dice, it also agreed perfectly on form. This is not the case with IA$_{bool}$, whose edit distance PRP was consistently lower than its PRP on Dice.

Results of pairwise t-tests showed that IA$_{part}$ performed significantly better than IA$_{bool}$ on both measures, in both datasets, as shown in Table 5.7. The magnitude of the effect indicated by the $t-$value is however smaller on $+$LOC.

The difference between the two algorithms is mainly due to the problems observed for IA$_{bool}$ in the previous chapter. As an example, (5.27) displays a human-produced formula (a), and the counterpart produced in the same domain by the Boolean algorithm (b). The partitioning algorithm outputs an identical description to the human gold standard in this case.

(5.27) (a) $\big[\langle\text{TYPE} : \textit{fan}\rangle \wedge \langle\text{COLOUR} : \textit{green}\rangle\big] \vee \big[\langle\text{COLOUR} : \textit{blue}\rangle \wedge \langle\text{TYPE} : \textit{sofa}\rangle\big]$

(b)
$$\begin{array}{c} \big[\langle\text{SIZE} : \textit{small}\rangle \wedge \langle\text{ORIENTATION} : \textit{front}\rangle\big] \\ \wedge \\ \big[\langle\text{TYPE} : \textit{sofa}\rangle \vee \langle\text{TYPE} : \textit{fan}\rangle\big] \wedge \big[\langle\text{COLOUR} : \textit{blue}\rangle \vee \langle\text{COLOUR} : \textit{green}\rangle\big] \end{array}$$



(a) Dice coefficient

(b) Levenshtein distance

Figure 5.3: Means Dice and Levenshtein scores of IA$_{part}$ and IA$_{bool}$

		+LOC			−LOC		
		Mean	**Mode**	PRP	**Mean**	**Mode**	PRP
IA$_{bool}$	**Dice**	.647	.667	1.4	.8	.8	4.3
	Edit	7.716	7	.7	8.335	7	3.5
IA$_{part}$	**Dice**	.7	.667	6.8	.88	1	44.7
	Edit	4.345	4	6.8	1.93	0	44.7

Table 5.6: Mean, modal and % perfect agreement on the two datasets

	+LOC ($t(147)$)	−LOC ($t(256)$)
Edit	9.279*	10.039*
Dice	3.787*	19.861*

Table 5.7: Pairwise t-tests comparing the two algorithms. (*$p < .001$)

The difference comes about because, although both algorithms have the same preference order (COS in this case), IA$_{bool}$ does not include COLOUR at first pass, since at this stage, only literals are being considered, and are only included if they are true of both referents, and remove some distractors. This is indeed the case for ORIENTATION and SIZE, which are therefore both included. However, this is a domain in which the minimal requirement to distinguish the referents is COLOUR. Therefore, IA$_{bool}$ begins to consider disjunctions of length 2, finding the disjunction of two COLOUR properties, and terminating. In contrast, IA$_{part}$ includes COLOUR immediately. The two values *green* and *blue* induce a partition, so that the description consists of two fragments, each consisting of the conjunction of TYPE and COLOUR, which are disjoined and semantically parallel.

This example represents an instance of the problems that IA$_{part}$ was designed to overcome. First, the logical form produced is identical to that derived from the human description, representing a partition that observes the Principle of Semantic Parallelism, and is transparently mappable to a Natural Language representation. Second, epistemic redundancy only occurs to the extent that two dissimilar referents have a shared property, or a highly preferred attribute is required for one disjunct, and is included in another to maintain parallel structure. It is for related reasons that the global evaluation in Chapter 4 found that the performance of the IA declined on plural domains. Note, in particular, that the modal score of IA$_{bool}$ on the plural data in the $-$Location dataset is .8, whereas its global mode on combined singulars and plurals was 1 on the $-$LOC data in Chapter 4 was 1.

In the above example, IA$_{bool}$ receives an edit distance score of 12, while the formula produced by IA$_{part}$ has a cost of 0. On descriptions containing locatives, performance declined, although more so in the case of IA$_{bool}$. A comparison of Figure 5.3(a) and 5.3(b) indicates two things. First, the distance between IA$_{bool}$ and IA$_{part}$ remains sharp on the $+$LOC dataset when edit distance is the measure of comparison. This suggests that as far as form is concerned, IA$_{part}$ benefits strongly from the partitioning strategy, though the edit score is still higher on $+$LOC than on $-$LOC. The distance between the two algorithms on Dice scores is not as sharp. This is mainly because the domains in the corpus were such that on a number of domains, the required attributes (which both algorithms had to include in order for the description to be distinguishing) included strongly preferred attributes, so that the two algorithms converged on content (though not on form, for the reasons outlined above).

5.5 Similarity, syntactic complexity and same-TYPE aggregation

As observed in §5.3.1, IA$_{part}$ will sometimes induce a partition on sets which is not motivated by the basic-level category of entities. To return to an earlier example, if R contains two same-type referents, say $\{e_4, e_5\}$ in Table 5.5, *makeTypes* will not partition this set, since they are both desks. However, on considering COLOUR, the algorithm will first consider *blue*, finding it to be the most discriminatory. Therefore, *updateDescription* is called with $R' = \{e_5\}$. This will satisfy the condition in [3.32], namely that $R' \cap R_{\text{DF}} \neq \emptyset$, and will result in a division of the set into two DFs. At the end of this process, the description will be the equivalent of *the blue desk . . . and the red desk.* This opens up the possibility of performing aggregation to return a description consisting of the equivalent of two coordinate adjective phrases and a single head noun, such as the *the red* *and blue desks*, ensuring that descriptions represent partitions induced by TYPE. This kind of aggregation is termed **Same-TYPE Aggregation**.

There are two questions that will be addressed. First, are there semantic constraints on this kind of aggregation, that is, can any pair of properties be felicitously disjoined within an NP? Second, are there observable limits on the complexity of a noun phrase which is the product of such an operation? Syntactically complex NPs might be very difficult to comprehend. Such issues have been raised in the past in the GRE literature, especially with respect to plurals (see especially Gardent, 2002; Horacek, 2004), but solutions have remained largely speculative, with no empirical grounding. Here, I shall present some empirical evidence using the British National Corpus (BNC)[8] and sketch one way in which this evidence might be used to improve IA_{part}. The focus is on coordinated adjectival premodifiers in plural NPs with a single, plural head noun (e.g. *the blue and red chairs*).

The TUNA corpus does not contain instances of reference to two entities with the same basic-level TYPE, so that there is no direct evidence for whether authors would partition a set if its elements had the same conceptual category. Hence, this analysis cannot rely on the semantic transparency of the corpus for evidence. This is one of the problems with the kind of study that gives rise to a dataset like the TUNA corpus: such studies tend to be highly labour-intensive because of their balanced and controlled nature. Hence, they also tend to be restrictive in the number of conditions that they represent in an experimental design.

Since the BNC data is not semantically transparent, some of the assumptions that underlie this analysis should be treated with caution. First, I will focus on definite descriptions in the BNC, though it is not possible to know whether these are always referential in the sense used in this thesis. Second, I will assume that syntactic constituents such as adjectives and nouns stand for 'properties', an assumption that could well be violated in certain cases. Despite the limitations inherent in these assumptions, the BNC is sufficiently balanced in terms of text genre to yield some general syntactic and semantic heuristics on how modifier phrases in plural NPs are realised.

Turning first to the semantic question, suppose there are two entities, both with the same value of TYPE, one of which, e_x, has the contrastive property $\langle$COLOUR : *blue*$\rangle$ and the other, e_y is described via the property $\langle$SIZE : *large*$\rangle$. Ignoring the parallelism constraint incorporated in IA_{part} (which would include at least COLOUR for e_y if it was available, having included it for e_x), the question is whether the two properties could be felicitously disjoined (or coordinated). Which of the following two descriptions would a person be most likely to produce?

(5.28) (a) the blue desk and the large desk

(b) the blue and large desks

Here is an intuition: A description such as (5.28b) is more likely to be perceived as ambiguous between a reading in which there is a blue desk and a large desk, and one in which there are n desks which are both blue and large. In the latter reading, the coordination of the two properties could be interpreted as the equivalent of a logical conjunction.

Coordination in natural language is ambiguous between the conjunction and disjunction reading. Hence, if the above intuition is correct, descriptions such as (5.28b) should presumably be

[8] http://www.natcorp.ox.ac.uk

avoided. The same intuition does not seem to arise in the case of *the blue and red desks*, perhaps because *blue and red* are sufficiently similar (they are values of the same attribute, among other things) to be assumed mutually disjoint. Another possible explanation for the intuition is that coordinating conceptually unrelated properties is odd because of independent restrictions on coordination itself: It has been claimed that coordination is semantically licensed when the coordinates are related or similar (Lang, 1984, as cited in Eschenbach et al., 1989). In short, the hypothesis offered is that *coordination of modifiers within a noun phrase is coordination of related properties*. If confirmed, this hypothesis would complement the principle of Similarity discussed earlier in relation to partitioned descriptions: partitioning *within* a noun phrase would be constrained by a comparable notion of similarity too.

The second question outlined is related to complexity. This issue is only partially dealt with here, since a full account of syntactic complexity would require an analysis of several different structures. My focus will be on the number of properties that can be disjoined in a single plural NP, and on how many adjectival constituents such NPs can contain.

5.5.1 Data

The BNC dataset consisted of noun phrases such as the ones shown below.[9]

(5.29) the front and rear metal panels (BNC:C/C9/C91:1109)

(5.30) the poorer African and Asian countries (BNC:C/C9/C94:713)

These are definite plural NPs, which consist of a single plural head noun. They have at least one premodifier adjective phrase (AP), consisting of at least two coordinated adjectives. Therefore, they roughly correspond to the kinds of 'aggregated' plural descriptions discussed in the previous sub-section. The data was collected using GSearch (Corley et al., 2001), a tool for searching through unparsed, morphosyntactically tagged corpora to find syntactic structures, based on a user-defined context-free grammar whose leaf nodes are corpus tags. GSearch was used to search through a sample of 1284 files in the written sub-corpus of the BNC. The grammar fragment used is shown in Figure 5.4.[10] The search targeted any noun phrases consisting of a morphologically plural head noun (indicated by the NN2 tag), and premodified by any number of APs, which were coordinated by *and* or a comma, or not coordinated. In addition, APs could themselves be modified by words such as *very* and *enough*. For example, the grammar considers *tall, dark and handsome* as a coordinated AP, while (5.30) above has *poorer* which precedes the coordinate *African and Asian*. To reduce the possibility of ambiguous parses, target phrases were NPs occurring as subject or object of a verb phrase.[11] Since the search results included NPs whose APs had no coordination, they were post-processed to find only those phrases conforming to the above description.

The search returned 1037 NPs. Of these, 13 turned out to be wrong parses, and were excluded from the sample. As examples (5.29) and (5.30) above indicate, there is a potential syntactic

[9] Here and throughout the following chapters, examples from publicly available corpora are cited in the format CORPUS:FILE:SENTENCE.

[10] A note on the grammar notation: $+$ and $*$ are quantifiers whose meaning is essentially the same as that in standard regular expression syntax. The notation $\langle \text{TAG} = A.* \rangle$ means 'any morphosyntactic tag starting with A (this covers the set of adjectival tags in the BNC, including those used to mark comparative and superlative forms. See http://www.natcorp.ox.ac.uk/docs/c5spec.html for a full description of the BNC tagset.

[11] Verb phrases were defined as a main verb and possibly one or more of the auxiliary *do*, *be* and *have*.

Terminal nodes (words and/or tags)
$N_{plur} \rightarrow \langle \text{TAG} = NN2 \rangle$
$A \rightarrow \langle \text{TAG} = A.^{*} \rangle$
$A_{mod} \rightarrow \{$ *enough, very, too* $\}$
Coord → *and*
Det → *the*

Non-terminals
$AP \rightarrow A^{+}$
$AP \rightarrow AP\ \ A_{mod}$
$AP \rightarrow A_{mod}\ \ AP$
$AP \rightarrow AP\ \ ,\ \ AP$
$AP \rightarrow AP\ \ Coord\ \ AP$
$NP \rightarrow Det\ \ AP\ \ N_{plur}$

Figure 5.4: Grammar used for GSearch query

ambiguity in the structure of the adjective phrases. In (5.30) for example, the phrase can be interpreted giving *poorer* a wide-scope reading, as in [poorer [African and Asian] countries]. This is indeed the parse obtained by applying the grammar rules in Figure 5.4. For the analysis, this was assumed to be the case, that is, an adjective phrase was considered coordinated if the adjectives were comma- or *and*- separated, wide-scope otherwise. Thus (5.31 contains a single wide-scope modifier, while (5.30 contains none.

(5.31) the [deeper [emotional, mental and spiritual]] levels (BNC:C/C9/C9V:1402)

(5.32) the [emotional, mental and spiritual] levels (BNC:C/C9/C9V:157)

5.5.2 Structure and complexity

For an estimate of the complexity limits on premodifier phrases, I used the following indicators:

1. The total number of premodifier AP constituents in the NP, whether these were coordinated or not. For example, (5.31) contains 4 premodifiers in total, while (5.32) has 3.

2. The number of coordinated adjectives in a coordinate adjective phrase. The coordinate AP in (5.31) contains 3 such coordinates.

3. The number of APs with wide scope over a coordinate AP. Example (5.31) has 1 of these (*deeper*).

Table 5.8 gives mean, mode (most frequent value) and standard deviations for each of these three indicators, averaging over all NPs in the dataset. It also gives frequencies and percentages of the minimum and maximum value of each, that is, the maximum number of wide-scope modifiers, the maximum number of coordinates in a coordinated adjective phrase, and so on.

Before summarising the salient points about the figures in the Table, it is worth pointing out a kind of construction that was conspicuously rare. This was the case where an NP contained two or more coordinate phrases. In principle, it is possible to have phrases such as *the blue and red, large and small chairs*. Only one phrase with this kind of structure was found in the sample. This is reproduced below.

		Mode	Mean	Standard Dev.	Min	Max
1.	AP constituents	2	2.2	.48	2	6
2.	coordinates	2	2.08	.303	2 (93%)	5 (.1%)
3.	wide-scope	0	.112	.353	0 (89.6%)	4 (.1%)

Table 5.8: Syntactic complexity figures from the BNC

(5.33) the impressionist and modern, contemporary and nineteenth century departments (BNC:E/EB/EBV:2141)

The figures in the Table can be summarised as follows:

1. Most definite NPs in the sample contain no more than two adjectival premodifiers in total, whether these are coordinated or not. Though more complex cases are attested (as witness (5.29) above), these are relatively rare.

2. Within a coordinate adjective phrase, assumed to correspond to a disjunction of properties, there are seldom more than 2 coordinates – this was the case 93% of the time. The most complex phrase found in the corpus consists of 5 coordinated adjectives. However, this occurred only once in the dataset, with 63 (6.2%) NPs having 3 coordinates and 7 (.7%) with 4. Examples of each are shown below.

   (5.34) the visual, auditory and tactile impressions (BNC:C/CL/CLP:961)

   (5.35) the physical, psychological, social and educational needs (BNC:C/CA/CAP:1956)

   (5.36) the intellectual, economic, scientific, technological and cultural achievements (BNC:E/EE/EE2:226)

3. Modifiers with wide scope over the NP, such as (5.29) above, were rare in conjunction with coordinated APs, with roughly 90% of descriptions having none.

To use these indicators in an aggregation algorithm, various possibilities suggest themselves. One is to set a threshold at the maximum complexity found. However, the rarity of the maximally complex NPs suggests that this is in general something to be avoided. Instead, I will use an average-case approach, which never aggregates two description fragments if the resulting NP exceeds a mean complexity value within a window of 2 standard deviations. Given that the mean number of adjectives within a coordinate AP was 2.08, this would suggest that no more than $(2.08 + 2SD \approx)$ 4 values should be disjoined in an NP. Thus, aggregation will return phrases like *the red, blue, green and brown chairs*, but nothing longer than that. Assuming that values of an attribute are disjoint, this will only occur in the kinds of domains discussed here when the algorithm is called with a set of referents of cardinality 4. As regards the number of disjunctions within an NP, the corpus data showed that, with the exception of (5.33), there is never more than 1 coordinate AP in a phrase.

This heuristic is combined with a further restriction on non-disjoined properties, corresponding to the wide-scope modifiers in the sample. At a mean of .112, the data suggests that in conjunction with a coordinate adjective phrase, the additional complexity obtained by adding such modifiers is restricted to $(.112 + 2SD \approx)$ 1.

A generalisation of these heuristics can be made. Let DF$_1$ and DF$_2$ be two fragments (corresponding to NPs) with the same TYPE. If IA$_{part}$ returns two such fragments, then $M_{\text{DF}_1} \neq M_{\text{DF}_2}$, that is, their modifier properties are not identical (if they were, partitioning would never have occurred). These are only aggregated into a new fragment if the resulting fragment has the maximal complexity defined below:

1. M does not contain more than one disjunction;

2. M does not contain more than one further property in addition to the disjunction.

5.5.3 Semantics

To obtain an indication of possible semantic constraints operating on adjectival coordination, I used three indicators. Two of these were measures of the semantic similarity of coordinated adjective pairs. The third was an indicator of whether the two adjectives were antonyms.

The first semantic similarity measure was distributional (DS), obtained from BNC corpus data. The measure was proposed by Lin (1998b), and is an information-theoretic measure that estimates the similarity of words based on their likelihood of occurrence in the same grammatical relations in a corpus.[12] For the purposes of this study, I used similarity estimates obtained using this measure by Kilgarriff (2003) and Kilgarriff et al. (2004), based on the BNC and incorporated in the SketchEngine database[13](see Lin, 1998a, and Chapter 6 for a description of the method used to obtain similarity estimates). SketchEngine contains a word similarity thesaurus for each of the three main parts of speech (noun, verb, adjective), in which the entry for a head word is accompanied by a list of same-category words, ordered by their similarity to the head word. The latter is computed using a large variety of grammatical relations. In this study, two words w_1 and w_2 had 0 similarity if w_1 was not found among the first 500 items in the thesaurus entry for w_2. A non-zero DS value indicates that two adjectives tend to occur in the same environments, that is, are 'used to talk about the same things' at least some of the time. Thus, *economic* and *political* are highly related because, among other things, they are frequently used as modifiers for the same nouns.

The DS measure returns a value in $(0, 1)$, where 1 indicates that two words occur in exactly the same grammatical contexts, and are therefore identical as far as this measure is concerned. In practice, values between 0.2 and 0.6 are considered indicators of high similarity . For example, intuitively highly related adjectives such as *economic* and *political*, or *good* and *bad* have values within this range. Since coordinate adjective phrases could contain more than two coordinates, I averaged the pairwise similarity of all pairs of adjectives in a phrase on the DS measure.

The second measure was an extension of a definition of similarity originally given by Lesk (1986), by Banerjee and Pedersen (2002). Lesk's measure is based on a comparison of the definitions or glosses of two words, comparing the textual overlap between them. The adapted version by Banerjee and Pedersen compares the glosses of word *senses* in WordNet, and defines an overlap as the longest common sequence of words occurring in two glosses. Apart from directly measuring glosses associated with the WordNet entry or synset of two senses, Banerjee and Pedersen's

[12] This measure is discussed in much greater detail in Chapter 6, as it features strongly in the empirical work presented there.

[13] `http://www.sketchengine.co.uk`

extension uses a variety of other WordNet relations, combining them into a single 'super-gloss'. For example, the gloss associated with the hyponyms of two senses can be used, and the hyponym gloss can also be compared to the synset gloss. For the present study, the array of relations used was restricted to the following:

1. the gloss associated with the synset;

2. the example of a usage of a sense given in its WordNet entry;

3. the 'see also' field associated with a synset, which offers pointers to related synsets in the WordNet Database.

The Lesk or GLOSS measure was selected because adjectives in WordNet are not organised in a taxonomy as nouns are; hence, a more straightforward ontological relatedness measure, comparing for example how conceptually distant two adjectives are, is not easily obtained. GLOSS for any sense pair was computed by combining the overlaps holding between entries of two senses in the above three fields. The score was normalised by the size of the glosses. Unlike DS, the measure, which returns a real value, does not have a ceiling, so that it is rather more difficult to interpret what a 'good' score is. The main indicator here will therefore be the proportion of coordinate adjective pairs for which the measure returns a value greater than 0. I also correlate this measure to DS.

Antonymy (ANT) was not a scale but an indicator variable, whose value indicated whether two adjectives were antonyms, that is, semantic opposites (e.g. *light/dark*, *fat/thin*). This took a value of 1 (`true`) if there was at least one WordNet sense of an adjective a_1 listed among the antonyms of one WordNet sense of its coordinate adjective a_2. This indicator is close in spirit to one interpretation of the hypothesis presented above, namely that two coordinate adjectives tend to be mutually exclusive values of the same attribute. However, it is only a partial approximation, because while antonyms could be considered 'values of the same attribute', the reverse is not necessarily true.

These three measures were used in this study to operationalise the notion of relatedness of coordinate adjectives, and to indicate whether the initial intuition was on the right track. Figure 5.5 displays the proportion of times a pair of coordinate adjectives was found to have a similarity value greater than 0 on DS and GLOSS, and the proportion of times a pair were found to be antonyms.

A fairly low proportion of coordinated antonyms is found in the sample, though at close to 20%, this is non-negligible. It is however lower than the proportion of pairs for which the DS and Lesk measures returned a value greater than 0. One reason for this is that WordNet is occasionally idiosyncratic in defining antonymy relations. Thus, *back* is considered an antonym of *front*, but *rear* isn't. The main reason, however, is that antonymy is only a partial indicator of relatedness, especially of whether adjectives are values of the same attribute. The other two measures of similarity indicate a high proportion of non-zero values. For the DS measure proposed by Lin (1998b,a), there were 29% of coordinates which had a similarity value of 0, representing those cases where one adjective was not found in the thesaurus entry for its coordinate. The corresponding figure for the Lesk measure is 37.8%. However, both measures had a higher mean score: 0.203 in the case of DS, and .11 in the case of Lesk. In the case of DS, 38.9% of coordinates

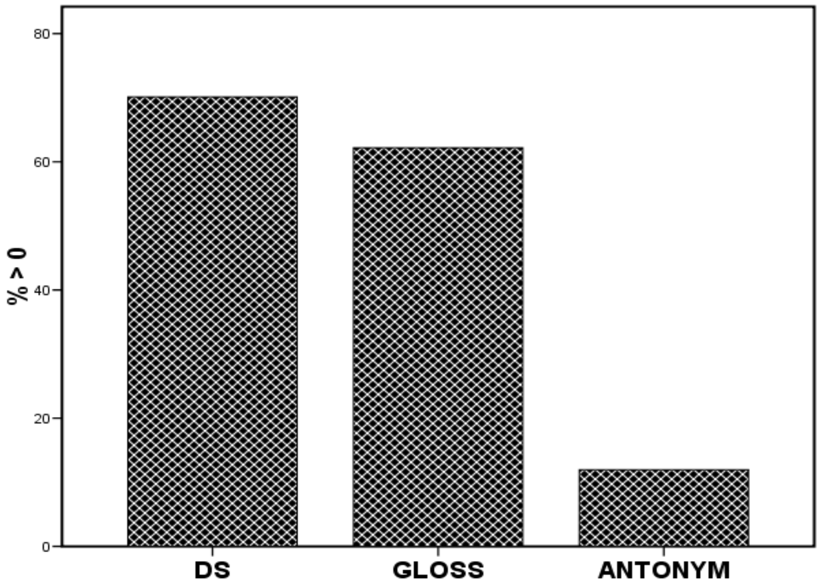


Figure 5.5: Proportion of coordinate adjective phrases with similarity > 0

had similarity values exceeding 0.25, with a score exceeding 0.4 in 17.7% of cases. The two measures were significantly positively correlated (Pearson's $r = .256, p < .0001$).

Overall, these results suggest that the semantic intuition stated at the outset is on the right track, and that coordinated modifiers – the rough equivalent of non-TYPE properties – are likely to be semantically related. In the present context, I will operationalise the notion of relatedness by stipulating that only properties which represent values of the same attributes should be disjoined. This is a rather simplistic interpretation of the findings, given that it does not necessarily correspond to the distributional and WordNet-based measures. In Chapter 7, I return briefly to this issue, and propose a revised interpretation of the same data.

5.5.4 Aggregation procedure

The extension of IA$_{part}$ with aggregation is based on the heuristics extracted above from the data analysis. The basic idea of the procedure is to aggregate only within a specific limit of complexity (no more than one wide-scope modifier, no more than four disjoined values of the same attribute). To enable this, the definition of a Description Fragment needs to be generalised. In Definition 5, fragments were stipulated as having a set M_{DF} of non-TYPE properties, corresponding to a logical conjunction. In the generalised definition, M_{DF} is a set of sets, and each element of M_{DF} is a disjunction. Thus, the modifier sets in DFs are now conjunctions of disjunctions of properties. The content determination procedure of Algorithm 3 remains the same, except that now, the algorithm does not add a property to M_{DF} in a DF, but a set containing that property at [3.35].

Definition 7. Description Fragment (Extended)

A Description Fragment (DF) is a triple $\langle R_{\text{DF}}, T_{\text{DF}}, M_{\text{DF}} \rangle$ where:

- $R_{\text{DF}} \subseteq R$ is the set of *intended referents* of the DF;
- $T_{\text{DF}} \in P_R = \langle A, V \rangle : A = \text{TYPE} \wedge R_{\text{DF}} \subseteq [\![\, T_{\text{DF}} \,]\!]$;
- $M_{\text{DF}} = \big\{ P \mid P \in \mathcal{P}(P_R) \wedge \bigcup_{p \in P} [\![\, p \,]\!] \cap [\![\, T_{\text{DF}} \,]\!] \neq \emptyset \big\}$

Algorithm 4 Aggregation algorithm

```
Require: $k$   ▷ max. length of non-singleton in $M_{\text{DF}}$
Require: $c$   ▷ max. no. of singletons in $M_{\text{DF}}$
Require: $d$   ▷ max. no. of non-singletons in $M_{\text{DF}}$
procedure aggregate (DF$_1$, DF$_2$)
    if $\big(T_{\text{DF}_1} \neq T_{\text{DF}_2}\big) \vee \big(att(M_{\text{DF}_1}) \neq att(M_{\text{DF}_2})\big)$ then   ▷ fail if DF$_1$ and DF$_2$ have different attributes
        return $\big(\text{DF}_1, \text{DF}_2\big)$
    else
        $M_{new} \leftarrow M_{\text{DF}_1} \cap M_{\text{DF}_2}$   ▷ initialise set with properties common to DF$_1$ and DF$_2$. $M_{new}$ contains only singletons
        $M_{\text{DF}_1} \leftarrow M_{\text{DF}_1} - M_{new}$
        $M_{\text{DF}_2} \leftarrow M_{\text{DF}_2} - M_{new}$   ▷ after update, $M_{\text{DF}_1}$ and $M_{\text{DF}_2}$ are of equal length
        if $\big(|M_{new}| > c\big) \vee \big(|M_{\text{DF}_1}| > d\big)$ then   ▷ fail if number of singletons or number of non-singletons is exceeded
            return $\big(\text{DF}_1, \text{DF}_2\big)$
        end if
        for $P \in M_{\text{DF}_1}$ do   ▷ compare each set in $M$DF$_1$ and $M_{\text{DF}_2}$ pairwise
            for $P' \in M_{\text{DF}_2}$ do
                if $att(P) = att(P')$ then   ▷ only aggregate values of the same attributes
                    $P_{new} \leftarrow P \cup P'$
                    if $|P_{new}| > k$ then   ▷ fail if max. length of disjunction exceeded
                        return $\big(\text{DF}_1, \text{DF}_2\big)$
                    end if
                    $M_{new} \leftarrow M_{new} \cup \big\{P_{new}\big\}$   ▷ update the new modifier set
                end if
            end for
        end for
    end if
    $R_{new} \leftarrow R_{\text{DF}_1} \cup R_{\text{DF}_2}$   ▷ new DF refers to $R_{\text{DF}_1}$ and $R_{\text{DF}_2}$
    return $\big\langle R_{new}, T_{\text{DF}_1}, M_{new}\big\rangle$
end procedure
```

- $[\![\ \text{DF}\ ]\!] = \bigcap_{P \in M_{\text{DF}}} \bigcup_{p \in P} [\![\ p\ ]\!] \cap [\![\ T_{\text{DF}}\ ]\!]$

The aggregation procedure compares pairs of DFs, which are candidates for aggregation only if they have the same TYPE. For example, the following two DFs could be aggregated.

(5.37) (a) ⟨⟨TYPE : *desk*⟩, { {⟨SIZE : *large*⟩} , {⟨COLOUR : *blue*⟩} }⟩

(b) ⟨⟨TYPE : *desk*⟩, {{ {⟨SIZE : *large*⟩} , {⟨COLOUR : *red*⟩} }⟩

Let $att(M_{\text{DF}})$ be the set of attributes represented in the set M_{DF} of some description fragment. Since disjoined properties within a fragment must be values of the same attribute, to aggregate two fragments, DF$_1$ and DF$_2$, it is required that $att(M_{\text{DF}_1}) = att(M_{\text{DF}_2})$. For example, *the large red chair* and *the blue chair* would not satisfy this requirement, and would not be aggregated, because this would allow the disjunction of the two COLOUR attributes, but would leave the SIZE attribute unaccounted for. By contrast, the two fragments in (5.37) would be merged to yield *the large blue and red desks.*

The aggregation procedure is shown in Algorithm 4. To take complexity limits into account, it requires three constants: k is the maximal number of disjoined properties in a disjunction [4.0]. This corresponds to the maximal number of elements of any set in M_{DF}, under the new definition of a fragment. c is the maximal number of wide-scope modifiers [4.0], which are not disjoined (the singletons in M_{DF}). d is the maximum number of disjunctions (the non-singletons in M_{DF}) [4.0].

In order for aggregation of two fragments DF$_1$, and DF$_2$ to take place, they must have the same value of TYPE; moreover, M_{DF_1} and M_{DF_2} must have exactly the same attributes [4.2].

The procedure then initialises a set M_{new} to contain those properties that DF$_1$ and DF$_2$ have in common [4.5]. These are non-disjunctive elements of the new DF. For example, in *the large blue chair* and *the large red chair*, *large* will be in M_{new} at this stage. Subsequently, a further check is performed. The remaining properties in M_{DF_1} and M_{DF_2} are known to be values of the same attributes. The algorithm fails and returns the original fragments if either one of the following conditions hold [4.8]:

1. M_{new} contains more than c elements. When this is the case, the number of wide-scope modifiers exceeds the limit c. This means that the returned fragment would contain several non-disjoined properties plus disjoined properties. This is necessarily the case, since no two DFs can have the same TYPE and exactly the same properties in M_{DF}.

2. $|M_{\text{DF}_1}|$ contains more than d elements. This corresponds to a situation where the two fragments, once aggregated, will result in a number of disjunctions that exceeds the threshold d.

Otherwise, the algorithm compares the elements of the set M_{DF} in the two fragments pairwise [4.11]. If any two elements P and P' contain values of the same attribute, they are merged into a set P_{new} [4.14]. A further check is carried out every time this happens, to see whether more than k properties are now disjoined, in which case, the procedure again terminates [4.15]. If not, M_{new} is updated with the new disjunction [4.18].

When this process has ended, the sets of intended referents in the two fragments are unified [4.23] and the result returned is a new fragment containing the TYPE and the disjoined properties.

This procedure therefore merges fragments within the empirically-motivated limits. For instance, the equivalent of *the large red desk* and *the small blue desk* would not be merged, as this would yield *the large and small, blue and red desks*, a construction not attested in the corpus data.[14].

This procedure has been integrated with the partitioning algorithm as a post-processing stage, following content determination, to merge any fragments that have the same TYPE in the description returned. An alternative would be to merge fragments on the fly, at the earliest possible opportunity. For example, during the loop in the update procedure of IA$_{part}$, which iterates through the description fragments, any fragment found to be complete (that is, it uniquely distinguished its intended referents) could be merged with any other complete fragments. To do this would however require that the main loop in *updateDescription* iterate through *all* fragments every time the procedure is called, rather than breaking as soon as R' is found to have been accounted for [3.27].

[14] Other possible constructions are possible, of course. For instance, the two example NPs could be merged into *the large red and small blue desks* Though this sounds intuitively less natural than *the large red desk and the small blue desk*, such possibilities would have to be further investigated on more corpus data.

5.6 Summary and outlook

This chapter began with an exploration of plural references in the TUNA corpus, motivated in part by the results of the evaluation in Chapter 4, which showed a dramatic decline in performance when the algorithms were extended to deal with disjunction. Based on considerations of both the form of plural referring expressions, and their content, a partitioning algorithm was proposed which has the following characteristics: (a) it partitions sets of referents opportunistically, breaking up the task of referring to a set into smaller sub-tasks; (b) it attempts to describe elements of the partition in similar ways, using the same attributes. The latter can sometimes yield overspecified descriptions, but this was found to be the case in the human-authored descriptions as well, modulo the codability of a property. This was included in the algorithm by incorporating a data driven regression model. The resulting algorithm evinced a markedly better match to the human data when compared to a procedure that was primarily motivated by considerations of logical completeness.

An extension of this algorithm to perform aggregation was also proposed. Though the data used to inform this procedure did not come from a semantically transparent corpus, it was found to complement the theoretical framework that gave rise to the content determination algorithm, insofar as a tendency to only coordinate similar modifiers was observed.

The empirical results and the behaviour of the algorithm highlight some further open questions. Though the data shows that overspecification is often desirable, a preference order, as incorporated in the IA, can make this excessive. For example, a reference to a large set, such as $\{e_4, e_5, e_6, e_7\}$ in Table 5.5 could result in a lengthy description containing COLOUR and ORIENTATION. On the other hand, a relatively dispreferred attribute, namely SIZE, would do the trick. As pointed out earlier, brevity-oriented algorithms, such as that proposed by Gardent (2002), would perform poorly on the data considered here. Nevertheless, the balance between using salient attributes and being concise remains unclear. For example, the extent to which a property is shared among referents may be involved in the decision to use an otherwise dispreferred attribute (e.g. $\{e_4, \ldots, e_7\}$ are all of the same SIZE), suggesting that the notion of similarity could be extended, and taken beyond a preference-order based strategy.

Another way of simplifying descriptions involves negation (*the desks which are not red*) (Horacek, 2004). Though it does not represent an insurmountable problem from a logical point of view (van Deemter, 2002), there are several untackled empirical issues in this area, especially in relation to possible semantic constraints on using negated literals. For example, descriptions such as *the things which are not tables* would presumably be ruled out in all but the most marked contexts. Moreover, do negated properties manifest the same codability or preference as their positive counterparts? For example, is $\overline{\langle \text{COLOUR} : \textit{red} \rangle}$ still preferred to $\langle \text{SIZE} : \textit{large} \rangle$? These questions must be left open to future research.

The next two chapters will build on the groundwork laid here, but will also move beyond the purely visual domains on which the investigation has so far been conducted. In particular, the focus now shifts to pluralities in discourse, where the properties that are used to describe them are not necessarily perceptual. As a result, the notion of similarity that was one of the motivating elements of this work is generalised to one of *semantic relatedness*. Chapter 6 begins with an empirical investigation into similarity factors affecting the way referents are categorised.

Chapter 6

Similarity and plurals in discourse: The Conceptual Coherence Hypothesis

> We are all familiar with the disconcerting effect of the proximity of extremes, or, quite simply, with the sudden vicinity of things that have no relation to each other . . . startling though their propinquity may be, it is nevertheless warranted by that *and* [. . .]
>
> MICHEL FOUCAULT, *The Order of Things* (1966)

6.1 Introduction

The account of plural reference proposed in Chapter 5 focused on perceptual principles underlying conceptualisation and formulation of plural descriptions, and similarity was defined in terms of perceptual attributes which were shared between elements of a target set and which were propagated across parts of a partitioned description. An 'adequate' plural description was therefore defined with reference to (a) its logical form, which reflected a partitioning based on an initial categorisation of the elements; (b) the extent to which the different elements of the resulting partition were described using the same attributes, modulo the codability of those attributes.

Here, I turn to the question of whether similarity operates at the level of conceptual categorisation, so that sets whose elements belong to different conceptual categories are mentally represented as a group (and license a plural reference) to the extent that a way exists of categorising and conceptualising them in similar ways. This can be considered an extension of the Category-Driven Reference principle to those situations where referents afford the speaker with multiple categorisations. Such a situation is the norm rather than the exception. For instance, persons that we know fall into different 'categories': one can speak about them in terms of their occupation (*plumber, professor*), their gender (*man, woman...*), their interests and achievements (*Nobel Prize winner*). Any number of such alternatives can apply to the same entity. The basic hypothesis I will test in this chapter is that, in referring to a set, categorisation of its elements must permit the hearer to infer something that those elements have in common. By hypothesis, talking about *the chemist, the physicist and the biologist* is better than talking about *the chemist and the tall blonde women* (assuming the two descriptions to be coextensive, so that both the physicist and the biologist are blonde women). Anticipating the terminology to be introduced later, the former description constitutes a more *conceptually coherent* cover of the set than the latter. A speaker who refers to these three entities using the first of these descriptions makes it easier for the hearer to form a unified, holistic representation of the set.

I will argue that the choice of how multiple entities are categorised depends on factors which are both local and global. Global factors include communicative intention (why a particular message is being formulated, what it is intended to convey, and so on). Such factors therefore go beyond the purely linguistic. Local factors are linguistic mechanisms: given our knowledge of language, of words and their context of use, we can infer from a speaker or author's lexical choice what different entities have in common, that is, what makes them *similar*. The focus of this and the following chapter will be primarily on local factors.

I begin (§6.2, p. 167) with a discussion of some further examples of plural reference in the next section, which will lead to the explicit formulation of a hypothesised constraint, called the *Local Coherence Constraint*, in part motivated by previous psycholinguistic work on plural reference. In §6.3 (p. 174) I turn to a discussion of why this hypothesis – if correct – has interesting implications for GRE. One practical problem that arises in this context is how to define the somewhat vague notion of similarity in a way that not only satisfies the intuitions motivating the hypothesis, but is also computationally useful. This is discussed in §6.4 (p. 176), where a number of different definitions of similarity are discussed, foremost among which is a *distributional* one, which defines similarity between words in terms of the likelihood of their usage in the same linguistic contexts. The different similarity measures are compared in three experiments reported in §6.5 (p. 178). These experiments are based on a phrasal/sentential judgement paradigm and used Magnitude Estimation to elicit from participants an estimate of how likely they judged a plural noun phrase to be used in some situation. In spite of the unprecedented nature of this judgement, validation data shows that subjects are remarkably self-consistent, and their judgements are strongly determined by similarity. Of the different definitions of similarity tested, it is the distributional definition that exerts the most significant influence. Two further experiments are then reported. The first (§6.6, p. 192) shows that given a choice, humans are very likely to refer to similarly-categorised entities together in a plurality. The second (§6.7, p. 196), placed participants in a situation where content determination – specifically, choice of categorisation – was required to distinguish a set of objects. In this experiment, semantic similarity is shown to be a determining factor in how people choose to conceptualise pluralities whose elements are not identical.

The hypothesis that these experiments test characterises a family of GRE models. Put simply, the aim of such models is to refer to entities by categorising them in ways that emphasises the similarity between them. This contrasts with the earlier family of GRE algorithms based on the Gricean maxims. Authors such as van Deemter (2002), Gardent (2002) and Horacek (2004) have all suggested, with different degrees of emphasis, that the adequacy of a plural description depends in large measure on its logical complexity and brevity. It is therefore interesting to ask to what extent this constraint interacts with the Local Coherence constraint, especially in those cases where they conflict. Such a case was exemplified above: the more coherent description involves a three-way partition of the set, and uses three properties (*physicist, biologist, chemist*), in contrast to the briefer *the physicist and the blonde women*. A final experiment in §6.8 (p. 200) compares these two models. Its results are surprising: while no preference is evinced for brevity models (in the specific way that Brevity is interpreted in the experiment), a strong preference is evinced for Conceptual Coherence. Though the experiment does not falsify a Gricean model which emphasises brevity, it also gives no evidence in its favour, while also showing that coherence has primacy in case of a brevity-coherence trade-off. At that stage, therefore, it is possible to consider some possible algorithmic incarnations of the family of algorithms characterised in this chapter.

6.2 Similarity and the status of plural referents in discourse

Consider the following examples of plural reference, obtained from the British National Corpus (BNC).

(6.1) (a) [The Smiths]$_i$ were not happy with the [Melody Maker]$_j$ piece and, not surprisingly, more court proceedings began to take place.

(b) The relationship between [[the band]$_i$ and [the paper]$_j$]$_{i+j}$ has never completely recovered.

(BNC: ART:2144-5)

(6.2) (a) Milo also wrote to the Austrian authorities and Metternich thought it advisable to remove [Vuk]$_i$ from Zemun, first to Buda and then to join [his]$_i$ Austrian wife and their two children in Vienna.

(b) Whilst in Vienna [he]$_i$ had the good fortune to meet [Petar II Petrovic-Njego]$_j$,, the Vladika of Montenegro who was passing through on his way to St Petersburg for his consecration as a bishop.

(c) The [[Montenegrin poet]$_i$ and the [Serbian philologist$_j$]]$_{i+j}$ made friends immediately [...]

(BNC:FSU:1490-1)

(6.3) (a) In June 1937, [Hoskyns]$_i$ died suddenly, at the age of only 52.

(b) [Ramsey]$_j$ felt the death to be a personal calamity.

(c) [[The master]$_i$ and [the pupil]$_j$]$_{i+j}$ had moved apart intellectually [. . .]

(BNC:A68:1350-2)

In all these examples, two referents are introduced using proper names (subscripted i and j in the examples). They are subsequently referred to in a plural NP ($i + j$), and this NP is a re-description: new properties are predicated of the referents. For example, the plural in (6.1b), refers to The Smiths and Melody Maker as *the band and the paper*. In this example, as well as in (6.2) and (6.3), there is the intuition that the referents i and j are re-described in a manner that highlights some relationship between them. Thus, Vuk and Petrovic-Njego are re-described as a *poet* and a *philologist* (6.2c), while Hoskyns and Ramsey are re-described as *master* and *pupil* (6.3c). The relationship between the two referents is partially inferrable from the properties that are predicated of them in the plural NP, an inference mediated by world knowledge (e.g. a master has pupils and/or followers; poets and philologists share an interest in literature and language by virtue of their profession).

The contrast between the singular NPs referring to i and j, and the description of $i + j$ in each example, suggests that there are two factors at work. First, there is a **shift in conceptual perspective** on the referents, from one which identifies i and j individually via proper names, to one which brings into focus a novel aspect of the entities. Second, the shift in perspective on the set is partially informed by the need to give a **coherence conceptual cover of the set** which is the object of the referential intention.

From a reader's point of view, the perspective taken in reference is identifiable from the author's lexical choice when categorising the referents. For the reader to be able to do this, he must be aware that such a choice is made from a number of possible alternatives, and the pragmatic effect of imparting a perspective on a referent or set hinges on the knowledge that such a choice is focusing some properties of the referents, but not others. For example, *the two men* might have done equally well to identify Ramsey and Hoskyns in (6.3), but the choice of words *master* and *pupil* suggests that the shift was not made gratuitously, but reflects an intention on the part of the author to highlight some relevant properties that relate the two referents. This relatedness and the change in descriptive material that brings it into focus allows the reader to infer the author's purpose in making the plural reference in the first place. The interaction of lexical choice and perspective-taking has been highlighted by Levelt (1999):

> [Lexical] choice is ultimately dependent on the perspective you decide to take on the referent for your interlocutor. Will it be more effective for me to refer to my sister as *my sister* or as *that lady* or as *the physicist*?
> (Levelt, 1999, p.226)

The claim that speakers have at their disposal several ways of categorising an object, and that their choice of categorisation, as reflected in their lexical choice, has the pragmatic effect of signalling a conceptual perspective on the objects to the listener, has been made by several authors. Lakoff and Johnson (1981), for instance, suggest that perspective-taking has the function of placing certain properties of an object in focus, while inhibiting the salience of other properties, because the focused properties help the speaker to achieve her communicative ends. That speakers often have several perspectives on a referent at their disposal is supported by studies of lexical entrainment in dialogue, in which a speaker's initial reference to an entity for which several possible categorisations exist, influences the way her interlocutor subsequently refers to it (e.g. Brennan and Clark, 1996). People have also been shown to be capable of forming novel categorisations of familiar artefacts, based for example on novel uses of these artefacts, such as using a boot to hold a door open (Barsalou, 1983). This results in a coexistence of multiple ways of categorising an object, based on its familiar and novel functions. It has been claimed that speakers will formulate references in which their lexical choice highlights a property which indicates which of many possible categorisations they have in mind (Clark, 1997a). For example, one might say *the doorstop* to refer to a boot if it has been used as a door-stopper. These insights are at the basis of H. Clark's Choice Principle (Clark, 1991) and the related Principle of Contrast of E. Clark (Clark, 1987, 1997a). These two principles hold that lexical choice is always made from a set of *non-equivalent alternatives* (even if the alternatives are extensionally equivalent). This non-equivalence rests on the prior assumption that true synonymy in language is nonexistent, an assumption that has been supported in philosophical and psychological work (e.g. Quine, 1953; Miller and Charles, 1991).

The variety of alternative perspectives available on a set of referents will depend on the knowledge that the speaker/author has of the entities referred to. However, choice is also restricted by context. For instance, in an academic context it might make sense to talk about my sister as *the physicist*, but the same contextual restrictions might preclude me from describing her as *the tall brunette with a brown handbag*, because it is less relevant to the discourse. For this reason, Kronfeld (1989) distinguishes between the *functional* relevance of a referential description, and its *conversational* relevance. Functional relevance refers to the by-now familiar notion of referential contrastiveness: if the intention is to identify a referent, then the properties used should be relevant to this intention (though of course other constraints apply; cf. §2.5, p. 33). On the other hand, a description is conversationally relevant if the properties used to identify a referent are licensed by the context. Kronfeld's example contrasts the two descriptions in (6.4).

(6.4) (a) New York needs more policemen

(b) The city with the world's largest Jewish population needs more policemen.

Assuming that the two alternative descriptions of New York City could have been uttered by the mayor in a public speech, the second of these might strike a listener as odd unless something in the context or situation (for example, a visit to New York by an Israeli diplomat) licensed the description of the city as *the city with the world's largest Jewish population*. Kronfeld's explanation of the mechanism at work here is formulated on Gricean communicative principles: on hearing (6.4b), a listener would need to make some assumptions as to why the mayor used that description instead of the more familiar (6.4a). This is required, according to Kronfeld, in order for the listener to infer what the speaker is 'up to', that is, what his intentions are. Since these intentions are not explicitly stated, but are inferred from the content of a description, they have the status of an implicature in the Gricean sense, and the inference process they trigger might cause the hearer to identify elements of the discourse context which might have influenced the mayor's choice of words. Therefore, while a speaker may intentionally choose a perspective on an object, context restricts the set of possible choices. A similar point is made by Aloni (2002), who argues that an appropriate answer to a question of the form *'Wh x?'* must conceptualise the different possible instantiations of *x* using a perspective or *conceptual cover* which is relevant given the hearer's information state and the context. For example, though the question *Who is the Prime Minister of Great Britain?* affords the speaker with many possible answers, whose content ranges from an exhaustive physical description to the man's most recent exploit on the international stage, only some of these answers will be relevant in a given context and for a particular interlocutor.

I will refer to contextual factors as **global constraints** on conceptual perspective and lexical choice in reference. The term 'global' here refers to the way longer-term communicative goals that a speaker might want to achieve are constrained over an extended period as a discourse unfolds.

Apart from these global constraints, however, the examples from the BNC also highlight **local constraints** on lexical choice in the production of plural referring expressions. The term 'local' refers to the dependency between parts of the same linguistic expression; hence, local constraints are operative at the stage where an intention to refer to a set is triggering a content-determination process for that set. To continue with Levelt's example, suppose I had two sisters, one of them a chemist and the other a physicist. Talking about them in the same context would probably preclude me from describing them as *the tall brunette with a brown handbag and the chemist*. The local constraint here is that the perspective taken on the referents, which is inferrable from the choice of lexical items, must be in some sense unified (conceptually coherent), so that the way one person is described influences the way another person is described. This is precisely what the corpus examples at the beginning of this section suggest; in each one there is the intuition that pairs like ⟨*poet, philologist*⟩ and ⟨*master, pupil*⟩ are in some sense similar and belong to the 'same perspective'. Where does this intuition come from? Unlike the domains that motivated the generalisation of Pechmann's Gestalts Hypothesis in Chapter 5, in the examples discussed so far it is not *perceptual* but *conceptual* similarity or relatedness that is at the basis of the intuition.

6.2.1 Types of local constraints

What I have called local constraints – operating within the NP – interact with global constraints and are related to reader expectations. The preceding examples suggest that conceptual categorisation of plural referents requires a plurality to be **conceptually coherent**, that is, the categorisation of i must be consistent with and similar to that of j in order for the plurality $i+j$ to be "licensed". Thus, there is a dependency between how one element of a plurality is categorised, and the description of other elements of the plurality.

The way referents are categorised also affects what other properties can be predicated of them because a categorisation brings to the fore salient aspects of an entity. For example, talking about Vuk as a Serbian poet in (6.2) brings into focus the person's occupation and, perhaps by some process of association, the fact that this kind of occupation bears some resemblance to other occupations (for instance, that of philologist). Because of this, properties predicated of a referent once it has been categorised (for instance, modifiers in NPs) may violate listener/reader expectations if they modify an aspect of the entity that the conceptual category does not make salient.

These restrictions have sometimes been referred to as restrictions of **conceptual combination** (Murphy, 1990; Clark, 1991). Theories of conceptual representation and lexical semantics (e.g. Murphy, 1990; Jackendoff, 1991; Pustejovsky, 1995) account for this phenomenon by proposing that nominals have structured lexical entries with slots or roles to which modifiers attach selectively, and noun-modifier composition foregrounds some aspect of the nominal semantics. Murphy (1990) showed that noun-modifier combinations such as *cold garbage* are difficult to process in spite of their being perfectly interpretable, as measured by reading time, compared to NPs where head nouns are combined with adjectives that are judged independently as more typical of them. The temperature of garbage is not one of its salient properties. Therefore, there is a dependency between the conceptual category to which an entity belongs, and the properties that can be predicated of that entity. In a similar vein, Cruse (1986) gives examples such as *spotless*, which is fine in combination with *kitchen*, but would be judged as odd in combination with *face*.

Further evidence for this was obtained in a study by Lapata et al. (1999), which found that the corpus-derived collocational probability of an adjective-noun combination was strongly correlated to their plausibility as judged by subjects. In an argument reminiscent of Shieber's (1993) discussion of Logical Form Equivalence, Lapata et al. suggest that if content determination is to distinguish between the shades of meaning of near-synonyms, and respect combinatorial restrictions, it needs to be lexically-driven, or at least lexically-informed.[1] Lexical knowledge extends the expressiveness of a generator in part because combinatorial restrictions are themselves clues as to the distinction between words that would otherwise be classed as synonymous (or nearly so).

[1] Similarly, Shieber argues that LFE is problematic only to the extent that strategic generation is 'blind' to linguistic realisation.

It seems likely that the full range of shades of meaning that distinguish lexical items cannot be captured by a theory of lexical semantics that does not take distributional regularities into account.

These constraints on noun-modifier combination have some bearing on the current discussion of pluralities as well. Kilgarriff (2003) gives the following example of an adjective modifying a coordinate (or disjunctive) plural NP.

(6.5) (a) old men and shoes

(b) old boots and shoes

The intuition is that the modifier *old* in (6.5b) modifies the entire coordinate NP (i.e. has wide-scope modification). The potential syntactic ambiguity is likely to go unnoticed. By contrast, the semantic dissimilarity between *men* and *shoes* in (6.5a) makes the ambiguity more evident. Chantree et al. (2005) showed that a distributional measure of similarity of nouns in a coordinate NP is a good predictor of modifier scope. Their interpretation is that *boots and shoes*, due to its high corpus frequency as a phrase, and the similarity of the head nouns, is a good candidate for a syntactic unit. My suggestions in the preceding section was that the similarity of conceptual relatedness of the nouns also makes the NP a good *conceptual* unit.

6.2.2 Psycholinguistic evidence for the role of similarity in plural reference

The idea that elements of a plurality must be conceptualised in similar ways has some currency in the psycholinguistic literature, and is related to the hypothesis that pluralities are groups with a separate status from their elements (i.e. they are gestalts). This was at the basis of an early paper on plural anaphora resolution by Eschenbach et al. (1989), who suggested that the discourse referent introduced by a plural anaphor (which they refer to as a Plural Reference Object or RefO) is licensed to the extent that its elements have a **Common Association Basis** (CAB), a term due to Lang (1984), who claimed it as a semantico-pragmatic principle underlying NP coordination. The authors' interpretation of this principle held that the possibility of "grouping [i.e. plural referent formation] depends on properties of the RefOs in question, namely, whether a CAB exists which constitutes a conceptual relation among the RefOs with respect to the situational parameters given" (Eschenbach et al., 1989, p.163). Subsequent experimental work has tended to focus on cases of pronominal anaphora with split antecedents (that is, anaphoric plural reference to entities introduced separately earlier in the discourse). The results of this body of work has converged on a few factors which make pluralities easier to process by readers, and more likely to be produced by authors:

1. *Similarity of discourse roles*: If referents in a discourse are described as participating in similar events or situations, this makes them better candidates for grouping into a plurality (Carreiras, 1997; Kaup et al., 2002). Thus, Kaup et al. (2002) found that a plural pronoun was resolved faster when its individual antecedents had had similar roles in a discourse. This is related to an earlier finding by Murphy (1984), who found that plural NPs whose earlier-introduced referents are *differentiated*, that is, have some property that distinguishes them, incur more effort in reading and resolution compared to NPs that did not differentiate their referents.

2. *Similarity of categorisation*: Koh and Clifton (2002) propose the *equivalence hypothesis*, which states that if a discourse entity is equivalent to another with respect to some property,

then the two can be grouped as a non-atomic discourse entity. The authors found evidence for this hypothesis using discourses in which a plural pronoun referred either to an *ontologically homogeneous* set, whose earlier-introduced elements belonged to the same broad ontological category (e.g. were all persons), or to an *ontologically heterogeneous* set. Interpretation was faster in the former case. A sentence continuation experiment showed a higher tendency for plural pronouns to be used as continuations in the homogeneous case.

3. *Structural symmetry*: A number of sentence continuation and online studies have focused on the way antecedents of a plural anaphor are introduced into the discourse. When they are introduced using a linguistic conjunction, if the method of conjoining makes the elements of the plurality symmetric arguments of a verbal predicate, then plural continuations to a discourse are more likely. Symmetric conjunctions include *and*, but also prepositional constructions in sentences of the form *X VP with Y*, where *with* results in an interpretation that assigns a common role to *X* and *Y* in relation to the *VP*. These constructions often trigger continuations using a plural anaphoric NP in sentence continuation studies, in contrast to asymmetric constructions such as *X VP for Y* (Hielscher and Müssler, 1990; Sanford and Lockhart, 1990; Moxey et al., 2004). Moxey et al. also replicated the effect in an eye-tracking study.

These studies provide convergent evidence for the idea that entities in a discourse model which have something in common – be it a common role with respect to a VP, or a common ontological category – are good candidates for a plural reference. Most of the experimental work has relied on a broadly-defined notion of ontological homogeneity of antecedents of a plural reference. Thus, Koh and Clifton (2002) used materials like (6.6), in which the three NPs are either all persons (HOMogeneous) or not (HETerogeneous).

(6.6) (a) A hurricane hit the city.

(b) A famous musician was missing.

(c) A distinguished novelist disappeared.

(d) { (HOM) A doctor also vanished.
(HET) A historic painting disappeared.

Koh and Clifton's equivalence hypothesis predicted that plural anaphoric references following a discourse in which homogeneous entities had been introduced would be likely to refer to all three entities. This was indeed found to be the case. However, ontological homogeneity as defined here is somewhat coarse-grained. It would not, for example, account for the intuition of relatedness in (6.1) between *band* and *paper*, without further elaborations about a metonymic shift whereby *paper* is understood to mean a group of people, as is a band. Even so, ontological relatedness does not easily capture the intuition that two individuals that belong to the same ontological category (e.g. are both people) might be categorised in a description in a way which incurs a mismatch between them. The description *the tall brunette with a handbag and the chemist* seems to violate expectations not because of ontological heterogeneity, but because it is difficult to see what describing someone as a brunette and someone else as a chemist is foregrounding in terms of the relatedness of the coordinated NPs.

In reviewing some of this work, Sanford and Moxey (1995) interpret it within the framework of *Scenario Mapping Theory* (Sanford and Garrod, 1981; Garrod and Sanford, 1982), in which the evidence for similarity (in the broadest sense of the term) as a facilitating factor in the representation and construction of pluralities in discourse is explained in terms of what interpretative possibilities are afforded by a text, given a reader's semantic knowledge. Thus, Sanford and Moxey (1995) view symmetry and the CAB as contributing factors to the comprehension of plural anaphors because they facilitate the process of mapping elements of a plurality to a common role which is either explicit in the discourse, or is cued from long-term semantic memory by some aspect of the discourse. More generally, their proposal is that when a plurality is introduced as part of an assertion, 'it is necessary that the assertion being made is coherent with what was already said, and that it maintains distinctions already made between individuals, or else motivates common role-mapping.' (Sanford and Moxey, 1995, p.31). In this proposal, therefore, the intuition of a common perspective on the pluralities $i + j$ in examples (6.1) to (6.3) would be explained as follows:

1. i and j are categorised and this triggers lexical choice;

2. $i + j$ is found to be **locally coherent** – that is, the plurality does not violate reader expectations – because the way the two elements are categorised cues semantic knowledge which includes a schematic representation of some scenario in which the two can be jointly mapped.

There is a slight difference of focus between the pragmatically-oriented theories of lexical contrast discussed above (Kronfeld, 1989; Clark, 1991, 1997a; Aloni, 2002), and the experimental work reviewed here. The former have as their primary focus a speaker's pragmatic intentions and how they inform lexical choice. Thus, H. Clark's Choice Principle is based on the Gricean notion of meaning as *m-intention* (Grice, 1957), whereby the meaning of an expression *x* "might as a first shot be equated with some statement or disjunction of statements about what 'people' [...] intend [...] to effect by *x*" (1957, p.66). Another possible view on the phenomena under discussion – closer in spirit to the experimental work discussed in this section – would place a greater emphasis on the "lower-level" processes of formulation and conceptualisation that lead to these kinds of lexical choices, given contextual parameters. I believe that these two views are not opposed; rather, they place different emphases, one on the intentional nature of linguistic communication, and the other on the mechanistic processes that make such communication possible.

6.2.3 The status of pluralities in discourse

At the basis of these hypotheses and the proposals by Eschenbach et al. (1989), is the further hypothesis that plurals in discourse are not represented as 'tokens', that is, as a set of disjoint individuals, as proposed in some early work (especially Johnson-Laird, 1983), but as holistic discourse entities. In line with the terminology used in the previous chapters, I will refer to this as the **Conceptual Gestalts view** of plural referent representation. A further terminological distinction will be useful: I use *plurality* to refer to the object of a referential intention, as distinct from a *set*, the mathematical construct whose elements are individuals in some domain.

The gestalts view has received some support from evidence of the so-called **Conjunction**

Cost, which is observed when a plural discourse referent has been introduced, and there is subsequent anaphoric reference to an element of the plurality (Garrod and Sanford, 1982; Gordon et al., 1999; Albrecht and Clifton, 1998; Moxey et al., 2004). Experimental evidence for the cost has usually been obtained from experimental materials in which a sentence introduces a plurality using a coordinate NP (e.g. *John and Mary*), with subsequent reference to one of the entities referred to (e.g. *John* or *he*). Compared to a plural anaphoric reference to the entire set (*they*), the singular reference takes longer to resolve, suggesting that the mental representation constructed by a hearer/reader is of the plurality as a whole, not its individual tokens. This effect has been found to interact with syntactic and semantic factors – including those listed above – affecting the discourse status of plural referents. For example, proper names seem to have a special status, in that when reference to an element of a conjunction occurs via a proper name, the cost disappears (Gordon et al., 1999). As argued by Sanford and Moxey (1995), this is probably due to the fact that proper names are semantically 'empty', that is, they function as rigid designators of an individual and usually have no lexical meaning beyond this function (Kripke, 1980).

The Conjunction Cost has led researchers to posit a role for plural NPs as triggers for the construction of a discourse referent in the mental model whose representation is as predicted by the gestalts view.

6.2.4 A hypothesis

I propose to extend the Principle of Similarity of the previous chapter to account for these phenomena in the following way. Rather than perceptual similarity, in these examples it is conceptual relatedness that characterises a set as a gestalt and enhances its status as a plurality in discourse. Thus, there is again an interaction between the Similarity Principle and the Principle of Category-driven reference. In the examples that motivated the preceding discussion, and in a substantial part of the literature on perspective and lexical choice, conceptual perspective on singular or plural referents was inferrable from the way entities were categorised. In describing a set, categorisation, which is a basic process in content determination, is constrained by the requirement that elements of a set be conceptualised in related ways. Hence, if a set is partitioned because of different category membership of its elements, the Similarity Principle predicts that the likelihood of a plural reference, and the ease with which it is comprehended, will depend on the availability of similar or related lexical items, corresponding to conceptual categorisations of those partitions, which give rise to a unified conceptual cover of the set. This proposal is incorporated in the following hypothesised constraint.

> **Local Conceptual Coherence Constraint on Plural Reference** (LCC)
> The process of referring to a set, and thereby introducing a plurality into a discourse, is facilitated if the elements of the set can be conceptualised in the same or related ways, using semantically similar properties that provide sufficient conceptual material to hold the elements of the plurality together.

6.3 The Local Conceptual Coherence Constraint and GRE

Before turning to the empirical evidence for the LCC, it is worth clarifying the nature of the challenges it poses for GRE. I will use the Knowledge Base in Table 6.1 as an example.

	TYPE	OCCUPATION	LEVEL	NATIONALITY	HEIGHT
e_1	man	postgraduate	third-year	maltese	medium
e_2	man	undergraduate	first-year	greek	tall
e_3	man	chef	–	italian	tall
e_4	man	engineer	–	french	medium
e_5	woman	research fellow	senior	scottish	medium
e_6	woman	research fellow	junior	chinese	short

Table 6.1: A simple knowledge base

Suppose a GRE algorithm were called with $R = \{e_1, e_2\}$ as input. Both referents can be distinguished on the basis of their NATIONALITY and OCCUPATION attributes. However, LCC predicts that describing elements of a set in dissimilar ways will result in the violation of expectations on the part of a reader, because it will be harder to infer a unified perspective on the set. In order for an algorithm to describe entities in a similar way, determining how to describe e_2 should take into account the description of e_1. For instance, a description like *the undergraduate and the maltese man* would, by hypothesis, violate listener expectations, since it is not obvious *a priori* what the two properties have in common. A description that categorised the two referents as *the postgraduate and the undergraduate* would probably satisfy expectations better. Note that describing entities in similar ways is not the same as describing them using the same attributes. For instance, a description of $\{e_3, e_4\}$ as *the chef and the engineer* – using the OCCUPATION attribute for either referent – could also, without a strong supporting context, result in expectation violation. The hypothesis would therefore predict that this description is worse than *the Italian and the Frenchman*, all other things being equal.

A strategy that took into account how an element of a set is described based on content selected for other elements is reminiscent of the strategy that sought to enhance semantic parallelism reported in the previous chapter, which was also motivated by considerations of similarity. The difference in the current examples is that using the same attributes to describe referents does not guarantee a satisfactory outcome. The LCC generalises this constraint, by changing the requirement to describe elements of a set with the same attributes, to a requirement of selecting the most coherent content for a description. Another limitation that the current example drops is the assumption that there is only one way of categorising an entity in a domain of reference. In Table 6.1, entities have more than one property that can be mapped to a head noun and function as the basic building block for the NP describing that entity.

Lexicalisation will also play a role if categorisation of entities restricts the choice of properties that can felicitously be predicated of them (as the evidence cited in §6.2.1 suggests). For instance, e_5 in Table 6.1 can be categorised in two ways, using *woman* and *research fellow*. Neither of these would be sufficient to distinguish the entity from e_6. Suppose OCCUPATION were selected; further properties to describe and distinguish e_5 include $\langle$NATIONALITY : *scottish*$\rangle$, $\langle$LEVEL : *senior*$\rangle$ and $\langle$HEIGHT : *medium*$\rangle$. However, the description *the research fellow of medium height* is intuitively worse than *the senior research fellow*.

Much of the work that has been used to motivate the LCC has focused on *lexical* choice. This is partially due to the fact that, from the reader's point of view at least, conceptual perspective is determined from the lexical structure of a description.

The LCC therefore poses a number of design challenges for content determination. Another challenge is to maintain the goal of GRE, which is to distinguish entities, without sacrificing completeness (i.e. without failing to distinguish referents for which a distinguishing description is available).

These challenges are taken up in Chapter 7. The rest of this chapter is dedicated to an empirical investigation of the predictions of LCC, focusing mainly on the conceptual coherence that results from categorising referents in similar or related ways. The experiments reported here also sought to find a computational definition of similarity for use in GRE. Therefore, they began with a series of experiments which first compared different similarity measures, and then sought to establish a causal connection between similarity and people's perceived likelihood of usage of a plural NP involving coordination. I therefore begin by describing the similarity measures tested.

6.4 Definitions of similarity

Intuitively, the similarity or relatedness of two words or concepts is a function of the things they have in common. Consider, for example, the two words *master* and *pupil* in (6.3c). To a native speaker, these two words are highly related. Like psycholinguistic definitions, computational definitions of similarity of words or concepts make different predictions. 'Having something in common' could be conceived in terms of ontological or taxonomic knowledge (Eschenbach et al., 1989; Koh and Clifton, 2002). Thus, in a taxonomy such as WordNet, at least some senses of *master* and *pupil* might be strongly linked. The fifth nominal sense of *master* in WordNet 2.1 is *schoolmaster*, while the first sense of *pupil* is that of a *learner*. Both are hyponyms of *person* or *individual*, and are therefore quite close in the taxonomy. An alternative take on the same intuition is incorporated in the view that similarity is deducible from *use* or context (Miller and Charles, 1991), a view whose correlate in the computational literature is a distributional (corpus-based) definition of similarity. This view is strongly related to the Firthian view of word meaning, and to that espoused by Wittgenstein (2001) in his later work.

In the first batch of experiments reported in the next section, the ontological view was initially operationalised using WordNet. Similarity under these definitions is similarity between *senses* of words ('concepts'), rather than words themselves. For the purposes of these experiments, the `WordNet::Similarity` Perl package developed by Pederson et al. (2004) was used to calculate similarity between pairs of nouns. Two of these measures augment the WordNet taxonomy with information-content or probabilities of use of word senses, which are provided by Pederson et al., and were based on the SemCor corpus, which is annotated with word sense information. The WordNet based measures were:

1. *WordNet Minimum Path*: This measure measures the multiplicative inverse of the length of the shortest path between two WordNet concepts a and b. Similarity is therefore defined in terms of the 'closeness' of two concepts in the taxonomy.

2. Resnik (1995): Augments the taxonomy with a monotonically increasing function $p : c \rightarrow [0, 1]$, where $p(c)$ is the probability of encountering an instance of concept c which subsumes a and b. $p(c)$ increases the further up the taxonomy c is, with a corresponding decrease in information content.

Possession	**VP: subject-of**	**VP: object-of**	**Noun: modifier**	**Adjective: modifier**
dissertation	attend	encourage	absentee	past
horizon	study	educate	school	old
knowledge	acquiesce	help	classroom	senior
need	interrogate	enable	grammar	junior
behaviour	teach	serve	geography	lower
absence	read	expose	biology	upper
skill	assemble	win	science	new
career	discuss	introduce	music	one-time
conduct	write	instruct	form	outstanding
book	explore	praise	ballet	individual

Figure 6.1: Common word sketches for *master* and *pupil*

3. Lin (1998b): Lin proposed an information-theoretic measure of similarity. Under this definition, the similarity of two arbitrary objects a and b is a function of the information gained by giving a joint description of a and b in terms of what they have in common, compared to describing a and b separately. Applied to a taxonomy like WordNet, this notion is formalised as the information content of the least common subsumer c of a and b, divided by the sum of the information content of a and b.

These three measures were compared to a distributional measure, which is an application of the definition given by Lin (1998b) to corpora (Lin, 1998a). This was already introduced in the previous chapter, for the study on adjectival aggregation (§5.5, p. 154), and is now discussed more fully. Applied to corpora to measure the similarity of words, the measure focuses on the grammatical relations in which two words occur (Lin, 1998a). Such relations are formalised as triples $\langle rel, w, w' \rangle$, where rel is a grammatical relation, w the word of interest and w' its co-argument in rel. For instance, some of the grammatical triples associated with both *master* and *pupil*, obtained from the British National Corpus, are shown in Figure 6.1, which restricts attention to five grammatical relations. Two triples for *master* represented in the figure are $\langle$*subject-of*, *master*, *attend*$\rangle$, and $\langle$*subject-of*, *master*, *write*$\rangle$. Both of these are also relations in which *pupil* is attested in the corpus, that is, this word is also found to be the subject of *attend* and *write*. However, the two words will not be attested in these contexts to the same extent; nor will they always occur with the same co-arguments in the same contexts. For example $\langle$*modifies*, *strict*, *master*$\rangle$ occurs reasonably frequently, but the corresponding triple for *pupil* ('strict pupil') is not (hence is not shown in the Figure). Therefore, prior to estimating similarity, the degree to which a target word w is associated with some co-argument w' in relation rel needs to be accounted for.

To quantify this, the measure of similarity takes into account the mutual information of w and w' in that relation, abbreviated as $I(rel, w, w')$. This expands on previous work by Church and Hanks (1990), who estimated mutual information by considering word co-occurrence probabilities in free text, within specific n-gram windows. This is equivalent to the measure of *category utility* by Gluck and Corter (1985), which has been used in machine learning approaches to conceptual clustering. In the present context, both measures could be described as estimating the degree to which knowledge about a word w' in rel decreases uncertainty about a word w. Mutual information is estimated as follows:

$$I(rel, w, w') = \log\left(\frac{\|\langle rel, *, *\rangle\| \times \|\langle rel, w, w'\rangle\|}{\|\langle rel, w, *\rangle\| \times \|\langle rel, *, w'\rangle\|}\right) \tag{6.7}$$

where $\|\langle rel, w, w'\rangle\|$ is the frequency of a triple and $*$ indicates any argument. The estimate of mutual information for w and w' in rel therefore takes into account (a) the overall frequency of the relation in question and (b) the overall frequency of w in that relation with w', scaling this by the frequency with which w and w' occur in that relation overall (Lin, 1998a; Kilgarriff and Tugwell, 2001). To estimate similarity between two words, we take into account their co-arguments in specific grammatical relations, weighted by their mutual information. Let $\sigma(w_1, w_2)$ denote the similarity estimate of two words w_1 and w_2, and let $F(w)$ be the set of words and relations which, together with w form an attested grammatical triple. For example, ⟨*subject-of*, *attend*⟩ is an element of both $F($*master*$)$ and $F($*student*$)$. Lin's formula to estimate similarity is as follows:

$$\sigma(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \tag{6.8}$$

Under this definition, the extent to which two words have common features is a function of the extent to which they are used in the same contexts, or talked about in the same way, rather than a function of their position in a taxonomic hierarchy or ontology. This definition emphasises language *use*, in line with the *contextual hypothesis* about word relatedness (Miller and Charles, 1991), which holds that the similarity of words can be estimated from the likelihood with which they can occur in the same contexts. This view has an intuitive relation to the Local Coherence Constraint introduced above. If the contexts in which words are used with significant frequency are grounded in *situations* in which things co-occur with significant frequency, then distributional semantic regularities may offer a window onto situational regularities. This view is in part supported by H. Clark's Choice Principle and E. Clark's Principle of Contrast, both of which assume that entries in the mental lexicon include situational parameters (Clark, 1991, 1997a).

As in §5.5 (p 154), the corpus-derived heuristics for estimating DS are obtained from SketchEngine[2] (Kilgarriff, 2003; Kilgarriff et al., 2004), which contains information about word similarity and the mutual information of grammatical triples, based on estimates from the British National Corpus (BNC)[3].

6.5 Eliciting ratings for disjunctive plurals

The first three experiments sought to substantiate the Local Conceptual Coherence hypothesis using a phrasal judgement design. Participants were shown phrases and/or sentences, of the form *the* n_1 *and the* n_2 *(VP)* and were asked to rate them in terms of their *perceived likelihood of usage in some situation*. The design manipulated the distributional similarity and/or the ontological relatedness of the two head nouns, based on the hypothesis that similarity will result in a higher perceived likelihood, as predicted by LCC. Apart from serving as a direct test of the hypothesis, these experiments also served the purpose of comparing the predictive power of the definitions of similarity introduced above.

[2] http://www.sketchengine.co.uk

[3] http://www.natcorp.ox.ac.uk

6.5.1 Magnitude estimation

The method used in these experiments, Magnitude Estimation (Stevens, 1957), was intended to investigate the systematic relationships that exist between perception of the magnitude of stimuli, such as brightness and amplitude, and their real physical magnitude. Participants are usually asked to rate the magnitude (say, the loudness) of a stimulus, such as a sound, by assigning it a number on a scale of their own choice. Initially, they rate a *modulus* item, to which all subsequent stimuli are compared. Crucial to the method is that participants' ratings maintain a constant scale throughout, and that they be proportional. Thus, if a modulus sound with perceived amplitude γ is rated 11.5 by a person, and the next sound is perceived as having amplitude 2γ, the person should rate the sound by assigning it a rating of 23. Scores are then normalised to enable comparison of perceived magnitudes among participants, by looking at the relative magnitude assigned to each experimental item in comparison to the modulus. Taking m to be the rating assigned by a person to a modulus item, and t to be the rating assigned to a subsequent stimulus, the normalised score t_n is calculated as follows:

$$t_n = \log\left(\frac{t}{m}\right) \tag{6.9}$$

This method has been applied successfully to linguistic judgements, ranging from ratings of the grammaticality of different sentences (e.g. Bard et al., 1996; Keller, 2003), to the acceptability of adjective-noun combinations (Lapata et al., 1999). It has several advantages over standard rating paradigms in which participants are asked to rate items on predefined (e.g. 5-point) scales, not least the fact that ratings are not restricted to an ordinal scale, but are placed on a ratio scale, where the difference between two judgements is meaningful. Another advantage is that a participant is free to select a scale of their own choice, thereby making it more likely that they will be aware of differences among gradations on the scale.

In psychophysics, subjective judgements of magnitude can be plotted against the real magnitude of the stimulus (e.g. the amplitude of a sound), so that people's accuracy of judgement and their self-consistency across different stimuli can be assessed. If people's judgements are proportional and accurately reflect real magnitude, the plot of real magnitudes against subjective judgements in log-log coordinates should fall on a straight line (that is, a regression line should have R^2 approaching 1, where R^2 is the proportion of variance in the data that the regression equation covers). Stevens (1957) showed how the relationship between subjective judgements of magnitude and real magnitudes can be systematically captured by a simple power relationship, where the perceived magnitude ψ is an exponential function kS^n of the real magnitude S, with constant k and exponent n.

An objective measure against which to compare subjective magnitudes is lacking in the present experiments, though some authors have proposed theory-internal methods of assessing subjective linguistic judgements against the 'objective' predictions of a theory (e.g. Keller, 2003). In the present case, plotting judgements of likelihood of usage of noun phrases against a corpus-derived or taxonomic measure of the relatedness of their constituents would be begging the question, since it is precisely the effect of the latter on the former that the experiments seek to investigate. Because of this, I used **Cross-Modality Matching**, a variant of the Magnitude Estimation task in which participants are asked to rate items in two modalities. For instance, they might be asked to rate a sentence using a numeric scale, but also by drawing a line whose length indicates their rating (cf. Bard et al., 1996). If participants are self-consistent in their judgements, normalised scores for one modality, regressed on the normalised scores for the other, should fall on a straight line with R^2 approaching 1. This is a way of ensuring the validity of the results, and is important in the present context because the kind of judgement task that people were asked to do – judging likelihood of usage of noun phrases, either alone or within sentences – does not have a precedent in the Magnitude Estimation literature.

6.5.2 Method

In all three experiments, participants were told that they would be seeing a number of phrases or sentences, and that their task was to rate them according to whether **they perceived them as likely to be used in some situation**. A sample of the instructions used for these experiments is shown in Appendix B, which also reproduces materials used in Experiments 2 and 3. Their primary aim was a direct test of the LCC, which predicts that disjunctive NPs will be more acceptable to speakers if their constituent nouns are similar.

Modulus and experimental items were judged in two modalities: numerically ('numeric judgements'), and visually. The latter involved moving a slider on a line ('line judgements'), such that the position of the slider reflected the positiveness of the rating. The slider position returned a real number in $(1, 100)$, though the actual numeric scale underlying the slider was unseen by participants. Based on previous work by Bard et al. (1996), participants were encouraged not to use zero values or negative numbers (as these would hinder normalisation to a logarithmic scale), and to avoid academic-style numeric scales, such as scales from 1 to 10, which have been found to limit people's ability to make fine-grained distinctions because of their limited range and their excessive familiarity.

The three experiments had roughly the same setup. All were conducted over the Internet. After reading the instructions, participants were shown a modulus phrase, which they were asked to rate on a numeric scale of their own choice, and also in a different modality, by moving a slider on a line so that the position of the slider indicated their rating. Once the modulus had been rated, subsequent phrases were rated (at different instances) both numerically and on the slider, and always in comparison to the modulus. To ensure that comparison was always taking place, the modulus, together with its rating in the relevant modality, was always shown together with the new item to be rated.

6.5.3 Experiment 1: a correlational study

The first experiment in the series used a correlational design, in which the ratings of disjunctive plural noun phrases given by participants were correlated to their semantic similarity, defined using the measures introduced above. The principal aim of the experiment was to obtain an initial indication of (a) whether the LCC hypothesis is on the right track, and (b) what definition of similarity correlated best with human judgements, in particular, whether a purely corpus-derived heuristic performs as well as, or better than, taxonomic measures. These questions were answered by correlating the mean normalised rating assigned by experimental participants to different noun phrases to the similarity of the noun pairs making them up, as defined by each measure.

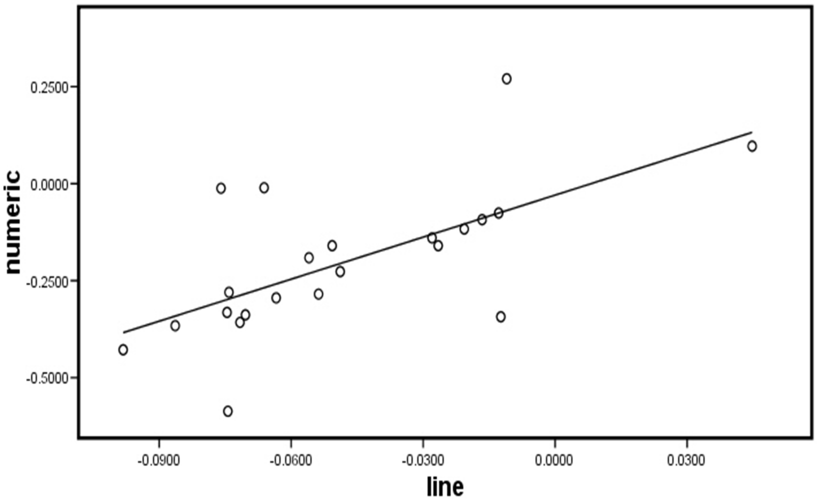


Figure 6.2: Cross-modality plot (log-log coordinates) for Experiment 1

In addition to the four measures of similarity, a random number in $(0, 1)$ was also used as a baseline test: a significant correlation to the random number would suggest that the intuitions underlying the LCC, as well as the corpus examples shown, are due to completely random factors.

Materials

For the construction of materials, nouns were extracted from the British National Corpus (BNC) and lemmatised using the morphological analyser described by Minnen et al. (2001). To avoid a confounding influence of word familiarity on people's ratings, nouns were divided into four frequency bands, ranging from **High** ($f > 500$ per million words) to **Low** ($f < 100$ per million). From each frequency band, 16 nouns were randomly selected and paired, to yield 8 pairs. From each pair, an NP of the form *the* n_1 *and the* n_2 was constructed. Thus, each NP contained head nouns that were roughly matched for frequency.

Participants and procedure

The experiment was carried out online at the University of Brighton. 63 self-reported native or fluent speakers of English, all University of Brighton staff, participated. After rating the modulus in the two modalities, they were exposed to successive randomised trials, each of which required a rating on *one* of the modalities. Thus, they saw each phrase twice, though at different instances.

Results and discussion

Figure 6.2 displays the cross-modality plot in log-log coordinates for the mean ratings given by each individual in the numeric and slider modalities. As the figure shows, there was some inconsistency on the part of a number of participants. A regression analysis revealed a value of $R^2 = .4$, which indicates high variance. Possible reasons for the inconsistencies are discussed below.

Correlations were generated between the mean rating of each description and the four similarity measures, as well as a random number in $(0, 1)$. Since some of the words used had more than one WordNet sense, the similarity between each pair of senses of the two words was calculated using the WordNet-based measures, and the highest one selected. The results are displayed in Table 6.2 below.

	MIN-PATH	RESNIK	LIN	DISTRIBUTIONAL	RANDOM
Pearson's r	.480	.535	.444	.576	.246
p	.05	$< .01$	$< .01$	$< .01$	ns

Table 6.2: Correlations to similarity measures in Experiment 1

As the table indicates, the least correlated measures were the WordNet Minimum Path and the Lin WordNet-based measure, whereas the highest correlations were obtained for the distributional Lin measure and the Resnik measure. Though significant, the correlation to the WordNet-based Lin measure was lower than that for the distributional version. Crucially, no correlation was found between judgements and the random measure.

The two similarity measures for which the highest correlations were obtained both use corpus data. In the case of Resnik (1995), $p(c)$ is calculated on the basis of sense frequencies in a sense-tagged corpus, while the similarity measure in (6.8) is entirely corpus-derived, and implicitly accounts for distinctions of word meaning as evinced by the occurrence of words in similar grammatical environments. Moreover, both the taxonomic measures that use corpus information correlated more highly than the purely taxonomic Minimum Path. These results suggest that speakers' intuitions seem to be influenced by a similarity metric which is, at least in part, determined by distributional information. This result, combined with the fact that people's judgements did not correlate to a random measure, suggests that the relevant measure of similarity is one which includes not only a strictly ontological view of taxonomic relatedness, but a view of relatedness that is based on the variety of relations that words occur in. The positive correlation to the measures also suggests that LCC is on the right track, and people judge plural NPs as more likely to be used when they involve similar categorisations of two entities.

Given that people in this experiment were rating perceived likelihood of usage, one might argue that the results support (though they do not show directly) that such a perception, based on an instruction that emphasises situational usage, is influenced by semantic similarity. However, this experiment only serves to point towards the 'right' measure of similarity. Because materials were randomly generated, there were some zero-values in the similarity estimates (between words that were completely unrelated). This tended to lower the correlation values, and created a lack of balance, since there were not equal numbers of high and low similarity pairs. In addition, the lack of self-consistency on the part of some individuals may have been due to the experimental task's not having been preceded by a *calibration phase*, in which participants are given some practise on the use of the different rating modalities, and which also serves to emphasise that ratings in either modality in relation to the modulus must be proportional. Furthermore, a correlational design still leaves open the question of whether there is indeed a causal link between people's estimates of the likelihood of usage of a disjunctive plural noun phrase, and the similarity of its constituent nouns. Experiments 2 and 3 were designed to overcome these problems.

6.5.4 Experiment 2: Similarity and ontological relatedness in disjunctive plurals

This experiment used the same method as the previous one, but in an experimental, rather than a correlational design. The aims of the experiment were to further investigate the effect of ontological similarity – operationally defined in terms of conceptual relatedness in WordNet – and the corpus-derived, distributional measure of similarity in (6.8), on the likelihood that people would

DS	OR	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

Figure 6.3: Materials used in Experiment 2

rate a disjunctive description as having a high likelihood of usage in some situation.

As in Experiment 1, the calculation of distributional similarity was based on the data in Sketch Engine. However, the list of grammatical triples used to calculate similarity by Kilgarriff et al. (2004) for this database includes NP coordination. The contribution of this relation to the overall similarity score is minimal, given the broad range of other grammatical relations. Moreover, supposing the similarity of $\langle w_1, w_2 \rangle$ were the object of enquiry, w_2 would occur at most once in the coordination triples of w_1, among several other words. Despite these facts, I recalculated similarity of the noun-pairs used to generate the materials of this experiment, in order to minimise the possibility of confounding factors. The set of triples on which similarity was calculated was the following:

1. *Subjecthood*: the likelihood of two nouns occurring as subjects of the same verb;

2. *Objecthood*: the likelihood of two nouns occurring as objects of the same verb;

3. *Modification*: the likelihood of two nouns being pre- or post-modified by the same adjectives.

Materials and design

Twelve pairs of nouns were manually selected from word lists generated from the BNC. From each pair, a disjunctive description of the form *the* N_1 *and the* N_2 was constructed. The materials represented all combinations of the following within-subjects factors:

1. **Frequency** (FR; 3 levels): Noun pairs were matched for frequency, which was either **High** ($f \geq 500$ per million), **Medium** ($500 > f \geq 300$ per million) or **Low** ($f \leq 100$ per million).

2. **Distributional Similarity** (DS; 2 levels): A pair of nouns n_1 and n_2 in a disjunctive description had **High** DS if n_1 was in the top 50 items in the Sketch Engine thesaurus entry for n_2, with $\sigma(n_1, n_2) \geq 0.2$ according to the new calculation. The pair had **Low** DS if n_1 was not among the top 300 nouns in the entry for n_2, and $\sigma(n_1, n_2) \leq 0.05$ according to the new similarity calculation.

3. **Ontological Relatedness** (OR; 2 levels): This was operationalised as the minimum path between two concepts in WordNet. **High** OR meant that the multiplicative inverse of the shortest path length between (the most highly related senses of) n_1 and n_2 was greater than or equal to 0.3. **Low** OR was defined as a minimum path value less than 0.01. As in Experiment 1, ontological relatedness was calculated as the highest value from all pairwise estimates of the senses of n_1 and n_2.

Some example phrases are shown in Figure 6.3; the full set of materials is given in Appendix B. As the examples show, it was possible to find pairs of words, such as *manager* and *council*, which had a high DS value, but did not have a high OR value. Nouns such as these belong to different ontological categories according to WordNet, whose IS-A taxonomy does not have a common root. For example, the three WordNet senses of *council* are hyponyms of *administrative unit*, *assembly* or *meeting*, all of which have *social group* as their least common subsumer. *Manager* is of course subsumed by *person*. The high DS value of these nouns is due to their tendency to occur in several similar contexts. For example, both *manager* and *council* are modified by *senior*, *general*, *technical*, and so on. In addition, logical metonymy is frequently found with group nouns such as *council*, so that the word stands in for its members in the context of a sentence. As this example, together with others in Figure 6.3 shows, there is quite a difference between a strictly taxonomy-based view of similarity, and one which takes into account the extent to which things are talked about in the same contexts.

Participants and procedure

27 self-reported native or fluent speakers of English did the experiment on the web. Once again, following initial rating of the modulus, the (2 modalities $\times$ 12 $=$) 24 trials were presented in randomised order, with each trial reminding the participant of the original modulus and the rating they had assigned it.

Calibration phase

Prior to commencing the rating task, participants went through a calibration phase, whose aim was to familiarise them with the concept of proportion and how it could be expressed in the two modalities. First, they were shown 4 numbers in order of magnitude. It was pointed out to them that the numbers formed a series, so that each number was larger than the previous by a factor of 3. They were then asked to move four sliders so that the position of each reflected the relative magnitude of one of the numbers in the series. The exercise was then repeated in the opposite format: participants were shown four numbers which formed a series, but whose order had been jumbled. They were also shown four sliders, each of which had already been placed in position. Their task this time was to select the number that matched each slider position.

Results and discussion

Figure 6.4 displays the regression plot of mean numeric and slider magnitudes for each trial in log-log coordinates. Regression indicated near-perfect self-consistency in ratings across modalities ($R^2 = .92$, $\beta = 0.96$, $p < 0.001$). This result shows that the task made sense to individuals, to the extent that they could give consistent judgements in two very different modalities. This lends additional validity to the results reported below.

A 3 (FR) $\times$ 2 (DS) $\times$ 2 (OR) ANOVA was conducted on the normalised ratings, using both participants (F_1) and items (F_2) as sources of variance. A significant main effect of DS was observed ($F_1(1, 26) = 47.909, p < 0.001$, $F_2(2, 11) = 53.505, p < 0.001$). The main effect of FR was also significant ($F_1(2, 26) = 16.083, p < 0.001$; $F_2(2, 11) = 7.272$, $p < .001$). No reliable main effect of OR was obtained ($F_1(1, 26) = 2.617, ns.$, $F_2(2, 11) = 1.081, p > 0.6$), but there was a reliable interaction of this variable with FR ($F_1(2, 26) = 9.414, p = .001$; $F_2(2, 11) = 3.472$, $p = .03$). The overall interaction of the three factors was also significant, but only by

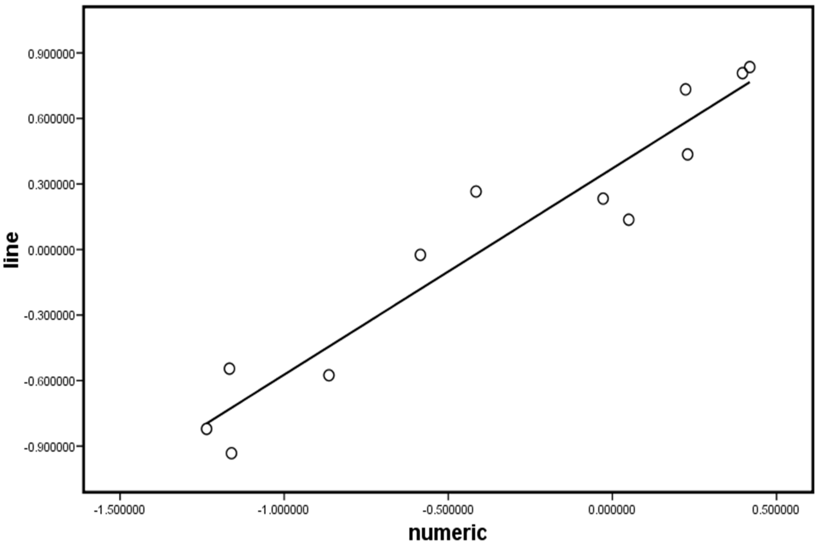


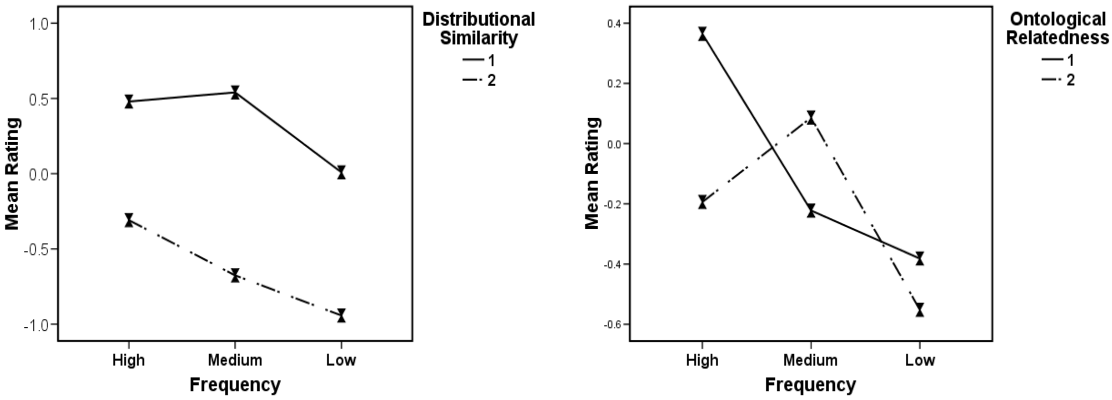
Figure 6.4: Cross-modality plot (log-log coordinates) for Experiment 2

subjects ($F_1(2,26) = 6.145$, $p = .004$; $F_2(2,11) = 2.278$, ns). No other interactions were significant.

Post hoc Tukey's comparisons of different levels of FR showed that the main effect was due exclusively to a difference between High and Low frequency levels (Tukey's $HSD = 3.558$, $p < .05$). Thus, participants' responses were sensitive only to large differences in word frequency. This also helps to explain the FR × OR interaction. As shown in Figure 6.5(a), High DS items were rated as more likely to be used than low DS items, at all levels of FR. By contrast, High OR items were only rated as more likely to be used when FR was very high or very low (Figure 6.5(b)).

The main effect of DS shows that people's judgement of disjunctive noun phrases is strongly determined by the extent to which the nominal constituents of those phrases tend to be used in the same linguistic contexts. This is in line with the Local Conceptual Coherence Constraint, and also supports the earlier result of a higher correlation with the distributional measure.

However, the lack of a main effect of Ontological Relatedness, and the lack of an interaction between OR and DS, is surprising, given previous psycholinguistic work which showed that ontologically homogeneous nouns tended to increase the likelihood of plural reference, and reduce the processing effort in reading (Koh and Clifton, 2002). The measure of distributional similarity used here will reflect ontological similarity to the extent that ontologically homogeneous entities are talked about in the same context. For example, although an NP such as *the greek and the postgraduate* violates the listener expectations predicted by the LCC, it still seems better than the description *the postgraduate and the table*. By contrast, in this experiment, word pairs which belong to different ontological categories were still judged as perfectly likely to be used, if they had high distributional similarity, but this may have been a result of OR having been defined in terms of the Minimum Path measure, which was found to be poorly, if significantly, correlated to people's judgements in Experiment 1. Moreover, ontologically unrelated words were often pairs consisting of an animate, human noun and a group noun that permitted an interpretation, via logical metonymy, that made it compatible with a 'human agent' reading. This would have overridden



(a) Distributional Similarity

(b) Ontological Relatedness

Figure 6.5: Mean ratings in different OR × FR, and DS × FR conditions in Experiment 2

a possible effect of ontological heterogeneity on judgements.

Apart from this possible objection, another potential problem with the phrasal judgement task in this experiment is that the judgement of phrases may evince a strong effect of DS, but this would disappear if the noun phrase were contextualised within a sentence. In other words, the effect of DS could conceivably disappear if a minimal amount of context were provided to support the two constituents of a disjunctive NP as co-arguments of the same sentential predicate. If the hypotheses made at the beginning of this chapter are correct, distributional similarity and/or ontological relatedness should still exert an influence on ratings in these cases, because similarity, calculated as a function of co-occurrence in the same grammatical environments, determines how easily two nouns can be 'mapped' to the same predicate, whether this predicate is verbal or adjectival. This possibility was addressed by the next experiment.

6.5.5 Experiment 3: Replication

This experiment attempted to replicate the results of Experiment 2 using a new set of materials, while also addressing its principal remaining concern regarding contextualisation and predicate mapping. To do this, half the participants rated phrases as before, but the other half rated the same phrases as subjects of sentences containing verb phrases with predicative adjectives, of the form *the* n_1 *and the* n_2 *were* AP. The main hypothesis tested in this Experiment was the same – namely, that high DS and OR would cause people to rate disjunctive NPs as more likely to be used.

Due to the problems observed above in relation to the Minimum Path measure of Ontological Relatedness, rather than relying on a hand-crafted taxonomy, the OR factor was this time manipulated in a more straightforward way, based on the distinction between animate, human entities on the one hand, and artefacts on the other.

A slight modification to the Cross-Modality Matching methodology was also made. It is possible that the high self-consistency across modalities observed in Experiment 2 was due to people having rated the *same* items twice, in different modalities. This does not reduce the validity of the data, since the cross-modal paradigm's main function is to demonstrate self-consistency, and the same item was rated twice in relation to the modulus, rather than to itself. However, rating the

DS	OR	**Example**
High	Animate	The secretary and the manager (were full-time)
Low	Animate	The technician and the nun (were good)
Low	Heterogeneous	The author and the novel (were popular)
Low	Heterogeneous	The politician and the shoes (were Italian)

Figure 6.6: Materials used for Experiment 3

same phrase twice may have introduced a bias and/or reinforcement effect, with people paying extra attention to the way they rated a phrase. More seriously, it may be that the effect found in Experiment 2 was restricted to the set of materials used there, and the overall self-consistency was an artifact of the NPs used. This possibility is worth doing away with. If the regressed line in the cross-modality plot can be found to approach perfect fit even when people have rated *different items representing the same experimental conditions*, rather than the same items, this would further validate the method, and make the results more reliable. Thus, in Experiment 3, participants never rated the same item twice.

Materials and design

A different set of materials was constructed in the same way as in Experiment 2, with the following differences. Frequency was not directly manipulated, while Ontological Relatedness was defined as a factor with three levels, as follows:

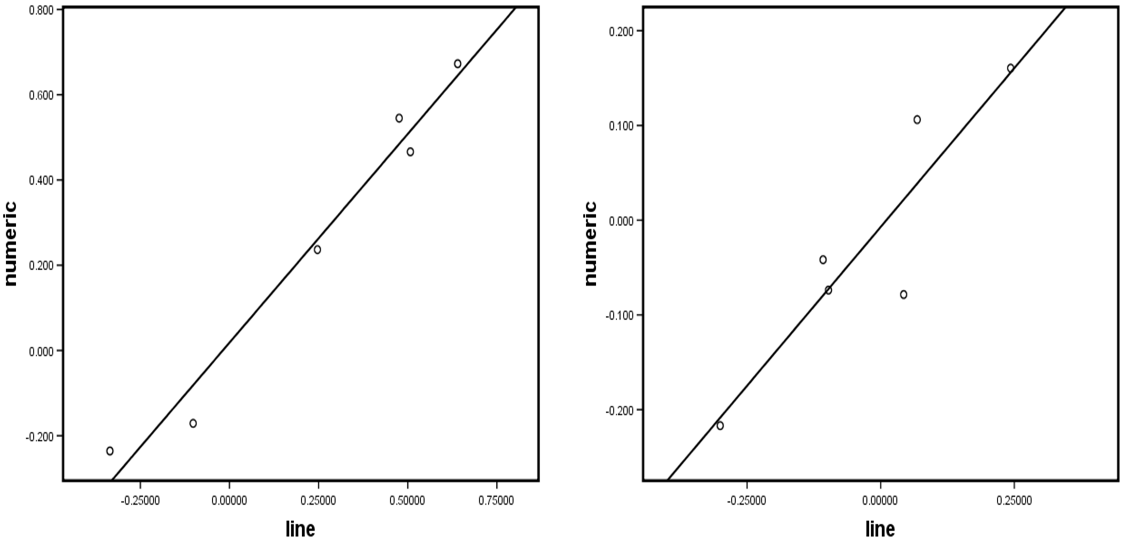
1. **Homogeneous, animate**: Both nouns in the NP denoted animate, human entities. Nouns were names of human roles, such as *plumber*, *author*, and *waitress*.

2. **Homogeneous, inanimate**: Both nouns in the NP denoted inanimate entities. These were always names of artefacts, such as *novel* and *table* and *desk*.

3. **Heterogeneous**: n_1 denoted an animate, human entity, while n_2 was an inanimate artifact.

12 NPs were constructed, so that each combination of the 2 (DS) $\times$ 3 (OR) within subjects design was represented twice. Six of these were designated as the *slider* or *line* trials, that is, they would be judged using the slider modality, while the other six were the *numeric* trials. Thus, rather than rate each NP twice, participants rated a different NP representing the same factorial combination in either modality. For counterbalancing purposes, two versions of the materials, A and B, were constructed, such that the six slider trials in A were numeric trials in B, and vice versa. 12 filler items were also included with the materials. These took the form of phrases or sentences involving singular NPs with coordinate adjective phrases.

There was an additional between-groups factor, **Stimulus Type** (ST), with two levels, **Phrasal** and **Sentential**. Half the participants saw NPs as they had done in Experiment 2, while the other half were exposed to the same phrases, but as subjects of a sentence containing a verb phrase. The VP always took the form of a copular verb, followed by an adjectival phrase in predicative position. These sentences always had the form *the n_1 and the n_2 were AP*. Adjectives for these phrases were selected based on BNC corpus data, so that an adjective was only used if it modified both nouns in the BNC with approximately the same frequency. Examples of the phrasal and sentential trials are shown in Figure 6.6.



(a) Phrasal condition

(b) Sentential condition

Figure 6.7: Cross-modality plots for Experiment 3, for each level of Stimulus Type

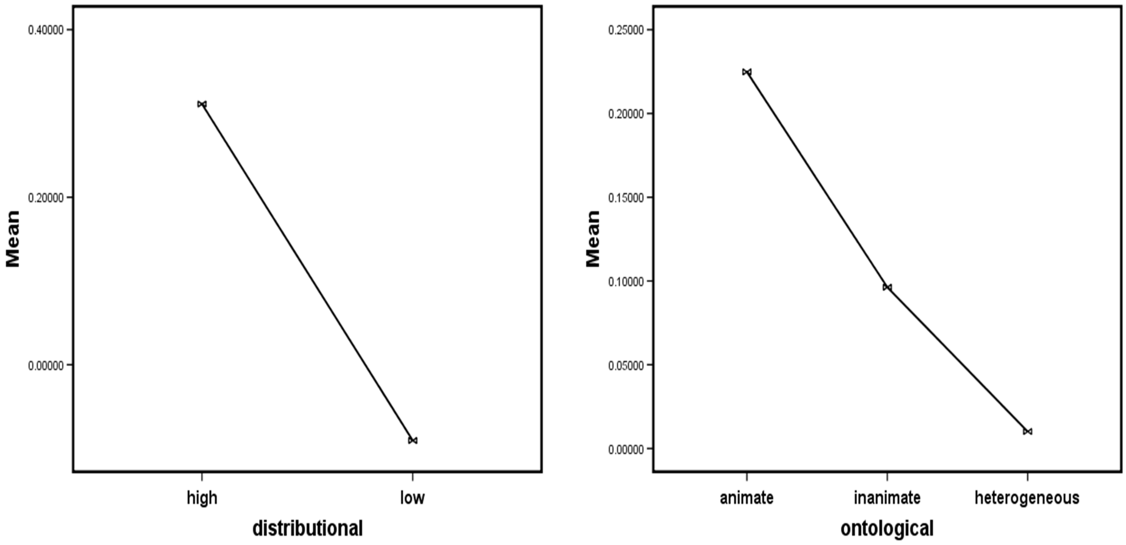
The predicative construction, rather than adjectival pre-modification, was used in order to avoid a potentially confounding interaction between noun-noun similarity and premodifier scope ambiguity. Kilgarriff's (2003) examples (see 6.5) and the evidence produced by Chantree et al. (2005), raise the possibility that when an adjective premodifies a coordinate NP whose constituent nouns have low distributional similarity, there is a greater likelihood that either only one of the constituents will be understood as being premodified, or the construction will be perceived as syntactically ambiguous. Either of these cases would confound judgements in the experiment, whereas the predicative construction with a plural copular verb made it clear that the adjective was to be applied to both nouns in the coordinate NP.

Participants and procedure

The procedure was identical to that of Experiment 2. Participants were randomly assigned to the Phrasal or Sentential groups. Within each group, half the participants saw the materials in Version A, while the other half saw them in Version B. The experiment was preceded by a calibration phase identical to that used in Experiment 2. A total of 147 self-reported native or fluent speakers of English completed the experiment. Of these, 39 were excluded from analysis, either because their ratings, once normalised, had a high (over 50%) number of zero values, indicating that the majority of their ratings consisted of scores of 1, or because they had made more than 50% errors in the calibration phase. This left 108 participants, 54 at each level of ST.

Results and discussion

Figure 6.7 displays the cross-modality plots for the judgements in the phrasal and sentential tasks. In both, participants once again proved highly self-consistent (Phrasal: $R^2 = .971$, $\beta = .985$, $p < .001$; Sentential: $R^2 = .829$, $\beta = .910$, $p = .01$). This was the case even though participants rated different items in different modalities representing the same factor combinations. However, there is a difference between the two conditions, in that people displayed slightly more variability in their cross-modal ratings of sentences than phrases, as indicated by a slightly higher variance



(a) Distributional Similarity

(b) Ontological Homogeneity

Figure 6.8: Ratings as a function of Distributional Similarity and Ontological Homogeneity

(a lower R^2 value). Thus, for a pair of different items, both corresponding to the same DS $\times$ OR combination, judgements were not always perfectly matched. Though the variance in the regression plot is low, its presence indicates some degree of uncertainty regarding ratings of an item, which was not present in the case of phrases.

For the by-subjects analysis, I conducted a 2 (DS) $\times 3$ (OR) repeated measures ANOVA, with ST as a between-groups factor. The by-items figures reported are the results of a 3 (OR) $\times 2$ (DS) $\times$ 2 (ST) univariate ANOVA on normalised judgements. In what follows, further exploratory data analysis is carried out via post hoc Tukey's tests.

There were significant main effects of DS ($F_1(1, 106) = 83.029$ $p < .001$; $F_2(1, 11) =$ 84.403, $p < .001$) and OR ($F_1(2, 106) = 15.348$, $p < .001$; $F_2(2, 11) = 8.623$, $p < .001$). Stimulus Type also exerted a main effect ($F_1(1, 106) = 11.818$, $p = .02$; $F_2(1, 11) = 40.415$, $p < .001$). This time, DS and OR interacted significantly by subjects ($F_1(2, 106) = 7.499$, $p =$ .001), but not by items ($F_2(2, 11) = 1.911$, $p > .1$). Stimulus Type also interacted reliably with DS ($F_1(1, 106) = 22.603$, $p < .001$; $F_2(1, 11) = 24.338$, $p < .001$) and with OR ($F_1(1, 106) =$ 17.609, $p < .001$; $F_2(1, 11) = 9.893$, $p < .001$). The three-way OR $\times$ DS $\times$ ST interaction was only significant by subjects ($F_1(2, 106) = 4.642$, $p = .01$; $F_2(2, 11) = 1.183$, $p > .3$).

The main effects of DS and OR are clarified in Figure 6.8. High DS items were consistently rated as more likely to be used in some situation than low DS items. As far as Ontological Relatedness is concerned, Animate NPs received higher ratings than Inanimates overall, but the lowest rating was for Heterogeneous items. To test which of the differences between levels of OR were meaningful, I ran a post hoc Tukey's test on the mean ratings at each level of the factor. Overall, the difference between Animate and Inanimate NPs turned out to be significant ($HSD = 3.279$, $p = .05$), as did the difference between Animate and Heterogeneous ($HSD = 5.465$, $p = .05$). However, there was no detectable difference between Inanimate and Heterogeneous NPs ($HSD = 2.186$, ns). This surprising effect is due to the interaction of DS and OR, shown in Figure 6.9.

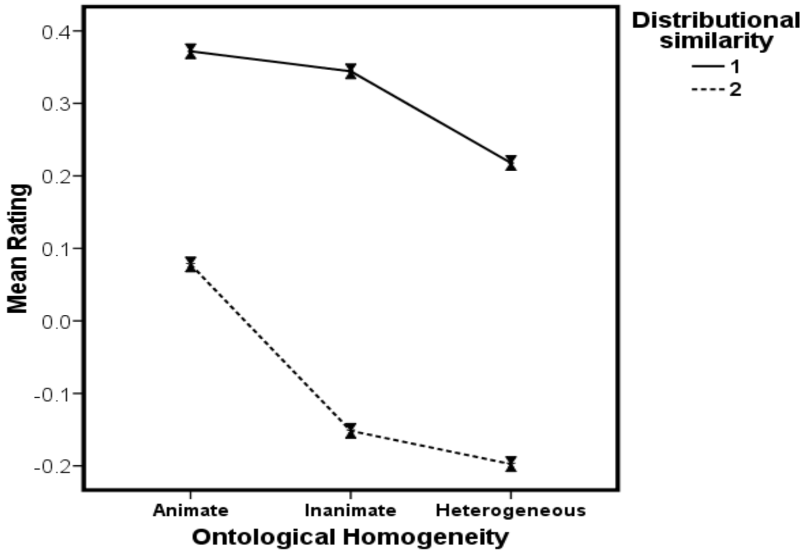
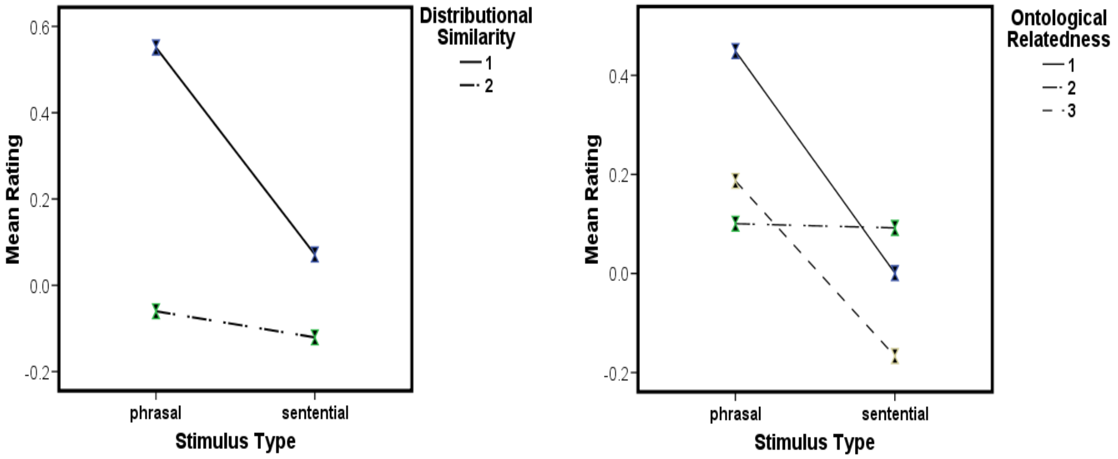


Figure 6.9: Distributional Similarity × Ontological Homogeneity interaction

For items with High DS, there was no difference between Animate and Inanimate noun-pairs ($HSD = .712$, ns), whereas the difference between them was significant when DS was Low ($HSD = 5.796$, $p = .05$). Similarly, both Animate and Inanimate NPs differed reliably from Heterogeneous pairs in the High DS condition ($HSD = 3.229$, $p = .05$), but only Animates differed from Heterogeneous pairs in the Low DS condition ($HSD = 6.966$, $p = .05$), whereas Inanimates were no different ($HSD = 1.169$, ns).

To summarise the results in relation to the main hypothesis in this experiment, the strongest impact on people's judgements is made by Distributional Similarity. When this is high, people rate highly those disjunctive NPs with ontologically homogeneous nouns, irrespective of whether the plurality denotes a set of animate, human entities, or inanimate artefacts. Ontologically heterogeneous NPs are consistently judged as much less likely to be used, but this effect is offset by DS. When the latter is high, even heterogeneous items are judged as quite likely. The situation changes when Distributional Similarity is low. Animates are still judged as better, but sets of inanimate entities this time cluster with heterogeneous sets: people rate them as far less likely to be talked about in the same NP, compared to Animates. Why should this distinction between the two kinds of ontologically homogeneous sets come about? I suggest it is because, while two distributionally dissimilar nouns are consistently perceived as unlikely to be talked about as a set, when both nouns denote animate, human entities, they have more features in common than when they denote artefacts. Artefacts could be loosely defined as non-natural kinds which have, as one of their main characteristics, a function (e.g. Pustejovsky, 1995), as well as a typical context of use. On the other hand, names of human roles or professions like *plumber* and *technician* not only carry information about roles and contexts, but also carry information about the fact that the denotata are human, though not all aspects of their semantics will be equally salient. To borrow a term from Sanford and Moxey (1995), the fact of being human may make it easier to conceive of a 'common scenario' for the entities denoted, even when they are dissimilar, whereas it is only when artefacts are similar (and have similar contexts of use and/or purpose), that such common



(a) ST × DS

(b) ST × OR

Figure 6.10: Phrasal and sentential ratings as a function of Distributional Similarity and Ontological Relatedness

scenarios can be imagined.

This experiment, in conjunction with the previous one, shows that statistical regularities in noun usage are a strong determinant of the perceived plausibility of a plural NP. The effect found when ontological considerations are taken into account (at least as measured here) is weaker, and exhibits a dependency on statistical similarity. Previous work on plural pronoun comprehension, which showed an effect of ontological relatedness, tended to focus on rather simple definitions of ontological categories. Moreover, proper names – which under one dominant semantic account are rigid designators with little if any lexical semantic content (Kripke, 1980) – often featured in these experiments. In the present work, the use of common nouns makes both the distributional and ontological relatedness of two NP constituents more complex to determine. The results suggest that the role that ontological homogeneity plays depends on the kinds of regularities that two nouns – denoting entities defined as homogeneous on some dimension – exhibit in everyday language use.

The next part of the analysis focuses on the difference between context-free NPs and the same NPs within the context of a sentence. Here, the main question phrased at the outset was whether the effect of DS/OR would disappear once a sentential context was provided. Given the significant main effects of both DS and OR, this was not the case. However, sentential contexts resulted in lower overall ratings for items compared to their phrasal counterparts. Mean ratings for phrases and sentences are shown in Figure 6.10, as a function of DS and OR.

The patterns of the interactions involving ST can be summarised as follows. High DS items, as well as Animate and Heterogeneous items, were rated lower by people in the Sentential, compared to the Phrasal, condition. The rating of Low DS items and Inanimates does not differ across conditions, whereas ontologically Heterogeneous pluralities, which received markedly lower ratings overall, were rated even lower in the Sentential compared to the Phrasal condition.

These results still support the main hypothesis, showing that similar noun pairs make for better descriptions also in a sentential context. The fact that sentences resulted in lower ratings is probably due to one main factor. Participants may have perceived the sentence structure used to be somewhat artificial. Since the instruction was to rate the perceived likelihood of usage, predicative constructions involving adjectives such as *Italian* and *good* may have been viewed as unlikely, despite the adjectives having been selected on the basis of their occurrence with both nouns in a corpus. Somewhat more puzzling is the lack of difference in the Inanimate case, between sentential and phrasal conditions. Once again, the same explanation as before could be given, namely, that inanimate entities have fewer salient features than animate, human entities, so that their (dis)similarity does not alter the rating of perceived likelihood dramatically enough.

6.5.6 Interim summary

The three experiments discussed above have shed light on the degree to which distributional similarity – defined in terms of regularities of occurrence in different grammatical environments, or on ontological relatedness – can be correlated, and causally related, to people's judgements of likelihood of usage of coordinate (disjunctive) noun phrases. So far, the results support the Conceptual Coherence Hypothesis. Overall, distributional similarity is the factor that has played the strongest role in the results obtained here. As hinted earlier, this may be because such a definition, comparing schemas which represent the usage of words in different contexts, takes into account the sort of knowledge, expectations, and situational regularities that affect our comprehension of discourse (Sanford and Moxey, 1999).

What these experiments do not show is whether the Conceptual Coherence Hypothesis, as a constraint on descriptions of pluralities, is a constraint on referential communication. Recall from the discussion in Chapter 2 that most definitions of the GRE task have focused on a purely extensional success criterion, only paying attention to coherence of representation of referents to the extent that such representations are approximated by preference orderings of properties in the Incremental Algorithm. If pure extensionality is indeed a feature of human reference, then the effects observed in these three experiments should be overridden – or at best be no more likely than pure chance – in 'real' referential situations, that is, situations where the properties of domain entities are known, and people are asked, not to rate phrases or sentences, but to produce them and identify entities in the process. The question therefore is: Can evidence for the Conceptual Coherence Hypothesis be found when the 'world' is explicitly represented in a domain? The next parts of this chapter attempt to reply to this question.

6.6 Experiment 4: Plural reference and aggregation

The purpose of this experiment was to investigate the effects of similarity on the likelihood of producing plural NPs. Specifically, it asked the following question: *Given a choice of referring to a set of three referents, two of them in a disjunctive (coordinate)* NP*, will similarity play a role in determining which two out of the three will be referred to together as a plurality?*

The experiment placed participants in a situation where they were buying objects from an online store. They were exposed to various trials, each consisting of a scenario where four pictures of objects were displayed with prices indicated for each. Three of these (the target referents) were identically priced, while a distractor object had a different price. Participants referred to the targets by completing a 2-sentence discourse:

S_1 The *object*$_1$ and the *object*$_2$ cost *amount.*

S_2 The *object*$_3$ also costs *amount.*

This discourse gave participants the possibility of referring to two out of the three objects in a plural NP in S_1. Given that only three objects in the scenario were priced with *amount*, it was clear from the context which of the three objects had to be referred to. If the effect of distributional similarity is also present in situations where the extension of a property is known, and is not overridden by the need to identify the entities, then participants should be more likely to refer to the two most similar entities out of the three referents in the plural NP in $S1$.

6.6.1 Materials and design

Eight trials or domains, each consisting of pictures of 3 targets and one distractor, were constructed, so that the targets were always identifiable as those which were identically priced. The pictures in a domain always represented artefacts. In the construction of trials, the principal factor manipulated was the distributional semantic similarity of the names of the three target referents, hereafter denoted $\{a, b, c\}$. Two of these, $\{a, b\}$ were the **designated targets**. The nouns referring to these pictures could be similar (High DS condition) or dissimilar (Low DS condition), but they were always dissimilar from the name of c. An example trial is shown in Figure 6.11, where the two designated targets are *clock* and *doll*. This trial represents an instance of the Low DS condition.

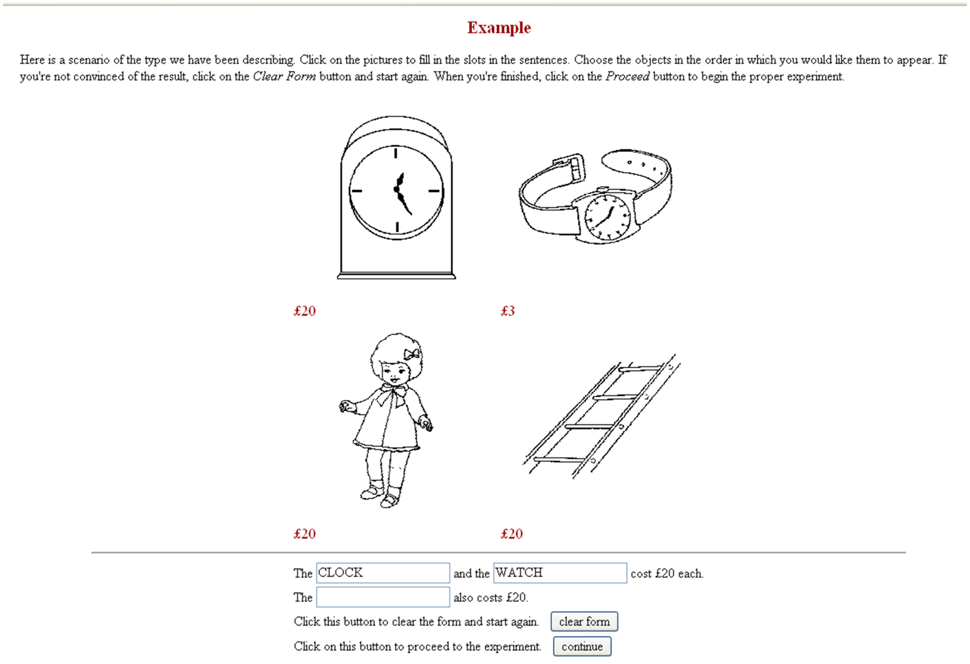
The two levels of DS were defined as in Experiments 2 and 3. Since these experiments showed that the effect of DS on judgements remained significant even when the similarity calculation omitted coordination from the set of relevant grammatical relations, this time I used the similarity estimates from the Sketch Engine thesaurus directly.

Pictures were selected from the Snodgrass and Vanderwart normed picture set, a set of drawings normed in a series of controlled picture-naming tasks with children and adults (Snodgrass and Vanderwart, 1980). For the construction of the trials, I used the norms collected among British English adult speakers by Barry et al. (1997), which contain the most frequent words given for a picture, each with an agreement factor (the proportion of people who gave the name for the picture in the norming study). All pictures selected had a most frequent name with an agreement factor of 85% or more. Semantic similarity of the designated targets was calculated for these frequent names.

To control for a possible bias due to visual similarity, an initial study was carried out in which participants rated the visual similarity of pairs of pictures on a ten-point scale. Each pair of pictures was rated by at least five individuals. They were asked to focus exclusively on visual properties of the stimuli (contours, straight lines, curves, etc), and ignore their impressions of the similarity in function of the objects represented. Based on the results of the picture-rating study, Visually Similar (VS) picture pairs (mean rating ≥ 6) were selected as designated targets for half the trials. The other half had Visually dissimilar (mean rating ≤ 2) picture pairs. Two sets of materials were constructed, for a total of (2 (DS) $\times 2$ (VS) $\times 2 =$) 8 trials.

6.6.2 Participants and procedure

The experiment was conducted over the Internet and completed by 27 self-reported native speakers of English. Trials were administered randomly for each participant. Each scenario or domain consisted of a 2×2 array of pictures, with the price of each picture clearly indicated beneath it.



Example

Here is a scenario of the type we have been describing. Click on the pictures to fill in the slots in the sentences. Choose the objects in the order in which you would like them to appear. If you're not convinced of the result, click on the *Clear Form* button and start again. When you're finished, click on the *Proceed* button to begin the proper experiment.

The CLOCK and the WATCH cost £20 each.

The ______ also costs £20.

Click this button to clear the form and start again.

Click on this button to proceed to the experiment.

Figure 6.11: Example domain for Experiment 4

The designated targets were never adjacent in the array. Beneath the domain was the two-sentence discourse to be completed. Participants completed the discourse by clicking on the pictures in the order in which they wanted the object names to appear in the sentences. Names of objects appeared automatically in the next available sentence slot when clicked. The option to reset the sentence slots and select different content was also provided. Participants could not type in the sentence slots.

The idea of asking participants to click on pictures was to cause participants to think of the name for a picture before using it, thus retrieving the lexical item they wished to use. If similarity plays a role in aggregating plural NPs, perhaps via a primitive priming mechanism, retrieving the lexical item for a picture should increase the likelihood that the next picture to be referred to in the same NP is the one most similar to the previous.

Pilot studies

Prior to running the experiment proper, two pilot studies were conducted. In the first ($N = 36$), participants were shown domains where pictures were placed in a single row, adjacent to each other. Because the results showed signs of a left-to-right clicking strategy on the part of some participants, this setup was abandoned in favour of one where pictures were shown in a 2×2 matrix, and the designated targets were never adjacent. In the second pilot study ($N = 48$), participants were asked to type the information missing in the descriptions. The pattern of results was identical to that reported below, and fewer than 8% of trials overall deviated from the lexical items predicted by the picture naming norms. Therefore, the picture-clicking methodology can be guaranteed not to result in a mismatch between participants' intentions and the actual descriptions produced, because the picture-naming norms are reliable.

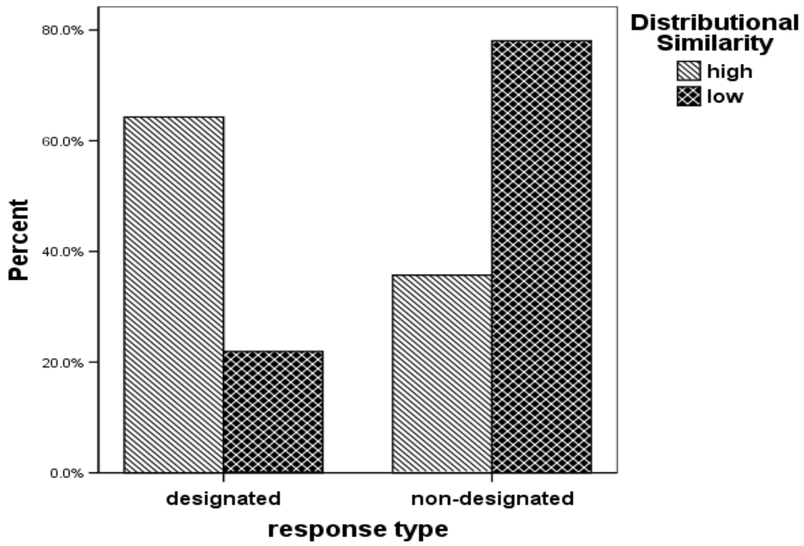


Figure 6.12: Proportions of response type per condition in Experiment 4

6.6.3 Results and discussion

Responses were coded according to whether the two entities in the plural description of Sentence 1 were the designated targets or not. Unsuccessful references were coded as errors.[4] Analysis is carried out on response proportions using pairwise Signed Rank Tests by participants (Z_1) and items (Z_2). I also report an initial χ^2 test on response frequencies.

In an initial analysis, the visual similarity of pictures turned out to play no role at all in people's selection of content, that is, people were no more likely to refer to the designated targets in S_1 when they were visually similar than when they were visually dissimilar. This is as expected, since Visual Similarity was only manipulated to control for potential biases. Results from different Visual Similarity conditions are combined in the analysis to follow.

As shown in Figure 6.12, participants referred to the designated targets 72% of the time in the High DS condition, compared to the 20.2% in the Low DS condition. Moreover, there were fewer plural references in S_1 consisting of one of the designated targets and the non-designated target, when DS was High. The difference in response frequencies across the two conditions was highly significant ($\chi^2 = 41.371, p < .001$). By participants, the proportion of designated responses was reliably higher in the High DS, compared to the Low DS condition ($Z_1 = 4.313, p < .001$), though it only approached significance by items ($Z_2 = 1.826$, $p = .06$). The same pattern was observed in comparing proportions of non-designated responses in the two conditions, with a significantly greater proportion of these in the Low DS condition ($Z_1 = 4.411, p < .001$; $Z_2 = 1.826, p = .06$).

Given the choice, participants prefer to describe similar entities in a plural description. Although the results showed that people referred to dissimilar entities roughly 30% of the time in the first sentence of a discourse overall, the trend is clearly and reliably in the predicted direction, with more references to the designated targets when they were similar. The reliability of these results is strengthened by two previous replications.

The main conclusion that can be drawn from this experiment is that participants show a

[4] These were usually cases where a participant erroneously selected the fourth distractor object in their plural NP.

strong preference for entities with similar types in plurals. This is predicted by the Conceptual Coherence Hypothesis, and suggests that distributional similarity at the lexical level is playing a role in determining people's choices. What the experiment does not address is the question of Content Determination. At the outset of this chapter, some motivating examples were given of discourses and referential domains in which it was clear that entities could be referred to in different ways and that by hypothesis, reference to plurals would be constrained by the availability of similar properties. This aspect of the LCC is perhaps the most crucial, since it has a direct bearing on the content determination strategy of a GRE algorithm that seeks to satisfy the Local Conceptual Coherence Constraint. Experiment 5 addressed this hypothesis directly.

6.7 Experiment 5: Content determination and distributional similarity

This experiment used a sentence continuation methodology to investigate the effect of distributional similarity on content determination. Participants were presented with domains containing multiple entities, in which more than one property could be predicated of an entity. Rather than using pictures, the domains were presented in the form of discourses, in which entities and their properties were introduced. The discourses served to place entities within a scenario or situation, in which they played a common role. By orthogonally manipulating the similarity of properties of pairs of entities, it was possible to test the hypothesis that the Local Conceptual Coherence Constraint is operative in content determination for plural reference.

The discourses sought to represent domains comparable to that shown earlier in Table 6.1, in which a given pair of referents, such as e_1 and e_2 in the table, can be referred to either using similar properties (*the postgraduate and the undergraduate*), or dissimilar properties (*the postgraduate and the greek*). These domains were presented discursively, rather than in a tabular or figurative display, in order to make the sentence continuation task more natural.

Apart from distributional similarity, the experiment also manipulated another factor, namely the kind of properties by which entities could be distinguished. The hypothesis in §6.2.4 focused on how entities are categorised, placing the burden of producing a conceptually coherent description of a set on the nouns used to categorise its elements. Modifiers (i.e. non-categorical properties) have been shown to be affected by the nouns they modify, as discussed in §6.2.1 (Cruse, 1986; Murphy, 1990; Lapata et al., 1999). This too was cited as a kind of local constraint on conceptualisation, but it was argued to follow from the expectations generated by a particular categorisation of an entity. Thus, it is less about describing a plurality coherently under a unified perspective, and more about describing entities in a way consistent with the properties that are made salient by the way they are categorised.

Nevertheless, there is the possibility that similarity of modifiers used to describe elements of a plurality affects coherence if the conceptual category of the referents is maintained constant. Thus, in this experiment the properties that could be used to identify the intended referents were either nouns or modifiers.

6.7.1 Materials and design

Sixteen discourses were constructed, with the same basic structure. An initial part consisting of one or two sentences introduced the scenario, or general topic, of the discourse. This introduced

> Three of the richest men in Europe were spotted last night dining at a London restaurant. All three are millionaires with a passion for fine arts and antiques.
>
> (e_1) One of the men, a Rumanian, is a dealer$_i$.
>
> (e_2) The second, a prince$_j$, is a collector$_i$.
>
> (e_3) The third, a duke$_j$, is a bachelor.
>
> The XXXXXXXXXX were both accompanied by servants, but the bachelor wasn't.

(a) Nominal condition

> Before selling his house, Dave decided to auction off some of the furniture. However, there were three vases he thought might be valuable, so he took them to an antique dealer for advice.
>
> (e_1) One of them was an Oriental$_i$ marble$_j$ vase.
>
> (e_2) Another one was a black vase, which was Persian$_i$.
>
> (e_3) There was also a bronze$_j$ vase. It was valuable.
>
> Dave decided not to sell the XXXXXXXXXX because he liked them both. He sold off the valuable vase for a lot of money.

(b) Modifier condition

Figure 6.13: Example discourses for Experiment 5

three discourse entities (e_1, e_2, e_3). Three subsequent sentences introduced two further properties for each entity. In half the discourses, the properties were nouns (Nominal condition; see Figure 6.13(a)); in the other half, they were adjectives (Modifier condition, see Figure 6.13(b)).

As shown in Figure 6.13(a), entities in the Nominal condition could be identified using different nouns (corresponding to values of TYPE in the earlier terminology of this thesis), while in the Modifier condition, they all had the same TYPE (*vase* in Figure 6.13(b)), but different modifiers. The properties were usually introduced using a predicative construction, as shown in the sentences in the Figure. In every discourse, two pairs could be identified using distributionally similar properties. These are indicated by subscripts in the examples. For instance, e_1 and e_2 in the Figure can be referred to as *the dealer and the collector*, and e_1, e_3 in Figure 6.13(b) could be referred to as *the marble vase and the bronze vase*. Crucially, however, they could also be described using dissimilar properties (*the bachelor and the collector*). High and Low Similarity was defined as in the previous experiments.

Each discourse was followed by one or two sentences which had a single, missing plural NP. This NP had to refer to two entities, something which was always indicated by the presence of the quantifier *both*. The target sentence contained another NP that referred to the third, non-target entity, in a construction that contrasted this entity to the missing target set. This NP was the subject of a subordinate clause beginning with *but* or *while*. For example, in Figure 6.13(a), the NP *the bachelor* is placed in the context of a subordinate clause beginning with *but*. The reference in this NP never contained a property similar to either of the target referents. For example, the second

NP in the continuation sentence in Figure 6.13(a) contains the word *bachelor*, rather than *duke*, which is similar to *prince*, and might bias content selection for e_2, one of the target referents. For counter-balancing purposes, two versions of each discourse were constructed so that the target referents in Version A and B were different pairs. This was done by changing the NP in the target sentence. The full set of materials is reproduced in Appendix C. Twelve filler items were also included. These consisted of discourses in which no more than 2 entities were introduced, and the continuation required singular reference.

6.7.2 Participants and procedure

18 native speakers of English, from the Aberdeen NLG Group database of experimental participants, completed the experiment. Items were presented in random order. Participants completed all 16 discourses, and were randomly assigned to Version A or B so that, for any discourse, there were roughly equal numbers of participants who referred to two different pairs of entities.

6.7.3 Results and discussion

Errors, consisting of references to a non-target entity, were omitted from analysis. The other responses were categorised as follows:

1. **Similar**: These were plural responses in which the two target referents were correctly identified using the similar properties provided in the discourse. There were three sub-categories of this response type:

   (a) *Disjunctive*: The plural reference consisted of a disjunctive NP with the two similar properties. E.g. *the duke and the prince* in Figure 6.13(a).

   (b) *Superordinate*: The plural reference consisted of a superordinate term that subsumed the two similar properties. E.g. *the noblemen*, where *noblemen* subsumes *prince and duke*.

   (c) *Include similar*: The two similar properties were used in a disjunctive NP, together with other properties. E.g. *the old, scratched car and the new, trendy one*.

2. **Dissimilar/other**: All other references were classified in this category.

Figure 6.14 shows the proportions of descriptions in each response category in the Nominal and Modifier conditions. Statistical results are reported comparing proportions of *Similar* responses overall (i.e. collapsing over response categories $1a$–$1c$), to *Dissimilar* (2) responses, and also comparing the *disjunctive* ($1a$) responses to Dissimilar responses. These comparisons are performed within Nominal and Modifier conditions. I also compare proportions of different response types across conditions.

Overall, *Similar* responses accounted for 66% of plural descriptions in the Nominal condition. Proportions of *Similar* descriptions overall (category $1a$–c) differed significantly from *Dissimilar* ($Z_1 = 2.719$, $p = .03$; $Z_2 = 1.997$, $p = .05$). Restricting attention only to those descriptions consisting of disjunctive NPs ($1a$) does not change the picture by participants ($Z_1 = 2.337$, $p = .01$), though the result is weaker by items ($Z_2 = 1.680$, $p = .09$).

The pattern of results is starkly different with the Modifier condition, where the difference between *Similar* responses overall ($1a$–c) and *Dissimilar* responses was neither significant by subjects nor by items ($Z_1 = .906, p > .3$; $Z_2 = .071, p > .9$). Focusing only on those disjunctive NPs

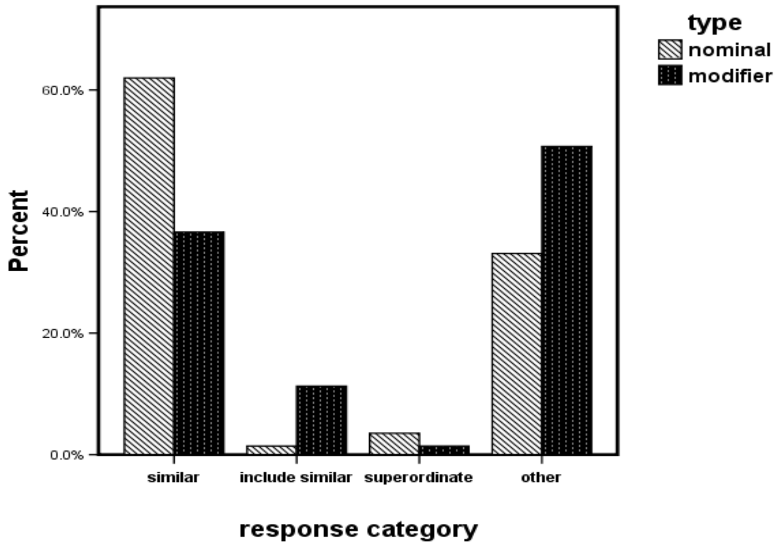


Figure 6.14: Response proportions in Experiment 5

with similar properties as defined in ($1a$) shows the opposite trend from the Nominal condition, with significantly more *Dissimilar* responses by subjects ($Z_1 = 2.126$, $p = .03$), though not by items ($Z_2 = .422, p > .6$).

A comparison of descriptions produced in the Nominal and the Modifier conditions confirms the trend shown in Figure 6.14: When all *Similar* responses ($1a$–c) are clustered together, the proportion is significantly greater in the Nominal condition, though only by participants ($Z_1 = 2.383, p = .02$; $Z_2 = 1.270, p > .2$). This holds even when attention is restricted to responses of category ($1a$) ($Z_1 = 3.237, p = .001$; $Z_2 = 1.612, p > .1$).

The results support the hypothesis that a constraint on similarity in the categorisation of elements of a set is operative in content determination for plural references. The hypothesis is not supported in the case of modifiers, where there was no detectable preference to use similar modifiers in constituents of coordinate NPs. Indeed, the opposite trend was observed in disjunctive NPs. This is a surprising result given that the corpus data in §5.5 (p. 154) indicated that similarity plays a role in adjectival coordination within NPs. However, those NPs consisted of plural descriptions with a single head noun. Therefore, it may be that modifier similarity constraints operate *within* NPs, but less so across coordinate phrases. A more likely explanation is that in the current experiment, participants made an effort to distinguish entities of the same type or category in the Modifier condition, and therefore maximised the variation or distinctiveness of the properties used to describe otherwise identical entities.

As suggested earlier, nouns are often thought by psychologists of as expressing TYPEs, which have a specific role in our mental ontologies, associated with *Gestalts*, whose primary function is that of categorising objects. If this is true then the findings can be interpreted as saying that a plurality is easier to represent mentally if the types on which it is based are similar than if they are dissimilar. As regards modifiers, while it is premature to suggest that LCC plays no role in modifier selection, it is likely that modifiers play a different role from nouns, namely to add information to an already-represented entity. When elements of a plurality have identical types (as in the modifier condition of the experiment), then perhaps the LCC is already satisfied, while selection of modifiers depends mainly on respecting the sorts of adjective-noun combination restrictions reported by other authors (cf. §6.2.1). In the algorithmic interpretation of the LCC which forms the topic of the next chapter, I will use the results on nouns to drive a content-determination strategy that attempts to maximise referent similarity. With respect to modifiers, I will tentatively assume that the relevant local constraint is that on noun-adjective collocational relatedness, while adjective-adjective similarity plays a role in aggregation of same-head NPs.

6.8 Experiment 6: Conceptual Coherence versus Brevity

As discussed in §6.3 (p. 174), the Local Conceptual Coherence constraint characterises a family of algorithms whose primary goal is to maximise the similarity with which referents are categorised in a description of a set. The experiments reported in the preceding sections have supported the LCC. However, the question arises as to how the LCC compares to the dominant, Gricean model that has informed most GRE algorithms to date. Since van Deemter (2002) proposed a logically complete version of the Incremental Algorithm, research on plural reference has often re-introduced Brevity as a desideratum, despite it being known, at least since Dale and Reiter (1995) – and also based on decades of psycholinguistic work – that the best distinguishing description is not necessarily the shortest one. The clearest example of the trend towards re-introducing Brevity as a constraint in GRE is perhaps Gardent (2002), but the same kinds of concerns can be traced in more recent work, such as Horacek (2004).

Brevity and overspecification have also been a focus of this thesis. The empirical work reported in Chapter 3, as well as the evaluation of Chapter 4, constitute falsifications of the Brevity-oriented model. Moreover, the partitioning algorithm of Chapter 5 actually introduces overspecification if the perceptual similarity of a set (and the parallelism of the partitioned description) is enhanced by doing so. How does Local Coherence fare in comparison to Brevity, especially when there is a potential trade-off between them?

In this experiment, participants were asked to compare pairs of descriptions of one and the same target set, selecting the one they found most natural. Each description could either be optimally brief or not ($\pm b$) and also either optimally coherent or not ($\pm c$). Optimal brevity here meant 'as brief as possible', while optimally coherent meant 'emphasising the similarity between categorisations of the intended referents'. 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 D*s, where the extension of *D* comprises the union of *B* and *C*). Since the following chapter will discuss specific incarnations of the LCC-based model and is partially motivated by the present results, it is worth spelling out the hypotheses tested by this experiment explicitly:

H1 $+c$ descriptions are preferred over $-c$.

H2 $(+c, -b)$ descriptions are preferred over ones that are $(-c, +b)$.

H3 $+b$ descriptions are preferred over $-b$.

Confirmation of H1 would be interpreted as evidence that, by taking coherence into account, an LCC-based algorithm would be on the right track. If H3 were confirmed, then earlier algorithms were (also) on the right track by taking brevity into account. Confirmation of H2 would suggest

Three old manuscripts were auctioned at Sotheby's.

e_1 One of them is a book, a biography of a composer.

e_2 The second, a sailor's journal, was published in the form of a pamphlet. It is a record of a voyage.

e_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.

Figure 6.15: Example domain in the evaluation

that, in references to sets, conceptual coherence is more important than brevity. Note that brevity here was defined in terms of the number of disjuncts in a disjunctive reference to a set; other kinds of brevity, including syntactic complexity of various kinds, were not taken into account (but cf. §5.5, p. 154).

6.8.1 Materials and design

Six discourses were constructed, each introducing three entities. Each set of three could be described using all 4 possible combinations of $\pm b \times \pm c$ (see Figure 6.15). Entities were *people* in two of the discourses, and *artefacts* of various kinds in the remainder. An example of the latter is shown in Figure 6.15. The full set of materials is reproduced in Appendix D. Properties of entities were introduced textually, as in the previous content determination experiment (see §6.7, p. 196); the order of presentation was randomised at runtime for each participant.

A forced-choice task was used. Each discourse was presented with two out of the four possible continuations. Each consisted of a sentence with a plural subject NP, and participants were asked to indicate the one they found most natural as a continuation. Thus, each participant made one selection for each of the six discourses. The six comparisons corresponded to six sub-conditions:

C1 **Coherence constant**

(a) $(+c, -b)$ vs. $(+c, +b)$

(b) $(-c, -b)$ vs. $(-c, +b)$

C2 **Brevity constant**

(a) $(+c, -b)$ vs. $(-c, -b)$

(b) $(+c, +b)$ vs. $(-c, +b)$

C3 **Tradeoff/control**

(a) $(+c, -b)$ vs. $(-c, +b)$

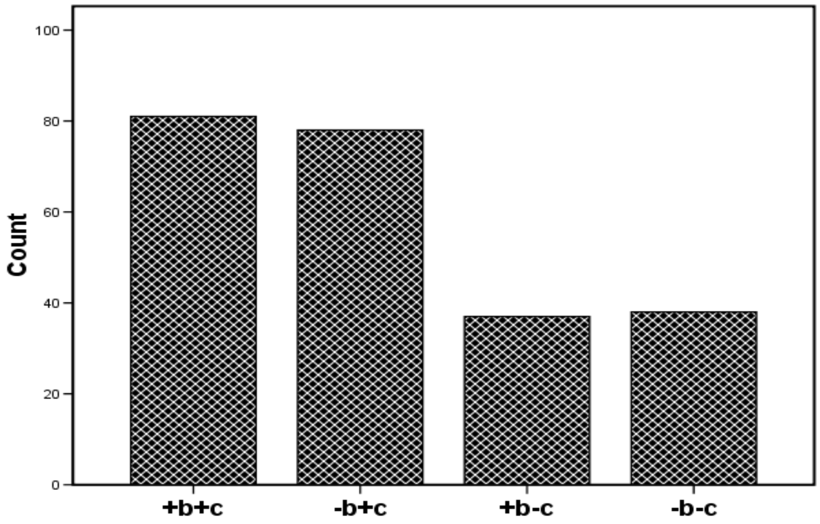


Figure 6.16: Frequency of selection of different coherence/brevity combinations

(b) $(-c, -b)$ vs. $(+c, +b)$

The first two sub-conditions tested H3 by keeping coherence constant and giving people a choice of whether they would refer using a brief or a non-brief description. The second set, which kept brevity constant, tested H2, by asking whether participants were more likely to use a coherent versus non-coherent description. The third sub-conditions are perhaps the most interesting. They involve cases where there was a trade-off between the two heuristics, so that the choice of a coherent description would trade off on brevity.

For counterbalancing purposes, a Latin square design was used. Six versions of each discourse were constructed. In each version, the comparison made reflected one of the above six sub-conditions. Participants were randomly assigned to one of six groups, so that participants in any two groups, while seeing exactly the same discourses, made the comparison on any sub-condition using different discourses.

6.8.2 Participants and procedure

39 self-reported native English speakers, all undergraduates at the University of Aberdeen, took part in the study, which they performed during a course practical. Discourses were shown in random order to each participant, and were presented through a web browser. The experiment was carried out in a computer laboratory at the University of Aberdeen.

6.8.3 Results

Results were coded according to whether a participant's choice was $\pm b$ and/or $\pm c$. Table 6.3 displays proportions of each response type within each sub-condition, where relevant. (Missing cells in the table are those where the relevant variable was kept constant in the sub-condition.) Figure 6.16 displays the proportion of times participants selected a particular type of description (i.e. a particular combination of the $(\pm b \pm c)$ factors). Note that each participant could select each option the same number of times.

The trends can be summarised as follows. Participants strongly preferred coherent descriptions, those predicted to be the most adequate by the LCC model. This is evident from the higher

	C1a	C1b	C2a	C2b	C3a	C3b
$+b$	51.3	43.6	$--$	$--$	30.8	76.9
$+c$	$--$	$--$	82.1	79.5	69.2	76.9

Table 6.3: Response proportions for each Evaluation sub-condition (%)

selection of $+c$ combinations. Surprisingly, there seems to be scarcely any effect of brevity: coherent descriptions were equally preferred whether or not they were brief, while descriptions that violated LCC, those in the $-c$ combinations, do not evince an impact of brevity. In other words, people's choices did not depend at all on whether descriptions were 'optimally brief'. To make these initial impressions precise, I report on tests comparing the overall impact of the different conditions, using a Friedman ANOVA by subjects (χ_1^2) and a χ^2 test on response frequencies by items (χ_2^2). To analyse the preference for brief versus non-brief, and coherent versus non-coherent descriptions, I report pairwise comparisons using a Signed Rank test by subjects (Z) and a χ^2 test by items, comparing proportions of responses in those conditions where people had a clear choice.

Overall, there was a significant main effect of condition, that is, proportions of response types differed reliably both by subjects ($\chi_1^2 = 107.3, p < .001$) and by items ($\chi_2^2 = 30.2, p < .001$). Pairwise comparisons between proportions of responses showed that there were significantly more $+c$ responses compared to $-c$, both by subjects ($Z = 4.682$, $p < .001$) and by items ($\chi^2 = 30.154$, $p < .001$). No difference was found between frequencies of $+b$ versus $-b$ descriptions, by subjects ($Z < 1, p > .9$) or items ($\chi^2 < 1, p > .8$).

To explore the data further, I compare response frequencies within individual conditions. In both conditions where coherence was kept constant (C1a and C1b), the likelihood of a response being $+b$ was no different from $-b$ (C1a: $\chi^2 = .023, p = .8$; C1b: $\chi^2 = .64, p = .4$). By contrast, conditions where brevity was kept constant (C2a and C2b) resulted in significantly higher proportions of $+c$ choices (C2a: $\chi^2 = 16.03, p < .001$; C2b: $\chi^2 = 13.56, p < .001$). No difference was observed between C2a and C2b ($\chi^2 = .08, p = .8$). In the trade off case (C3a), participants were much more likely to select a $+c$ description than a $+b$ one ($\chi^2 = 39.0, p < .001$); a majority opted for the $(+b, +c)$ description in the control case ($\chi^2 = 39.0, p < .001$).

The results strongly support H1 and H2, since participants' choices are impacted by Coherence. They do not indicate a preference for brief descriptions. This might be seen as echoing Jordan's finding Jordan (2000b,a), to the effect that speakers often relinquish brevity in favour of observing task or discourse constraints. It also supports the earlier findings reported in this Chapter and in Chapter 5, where similarity and codability were the main forces affecting people's content determination decisions, often resulting in overspecification of a particular kind. It seems, however, remarkable that the experiment shows up no brevity effect in situations where it is unclear that any purpose was served by being non-brief (for example, in the case where both descriptions were coherent, and differed in length). It seems that in such conditions, participants made a choice purely on a chance basis, even though no trade-off was present. Were speakers concerned with brevity, they would be expected to opt for the $+b$ descriptions.

Since this experiment compared the LCC model against the current state of the art in references to sets, these results do not necessarily warrant the affirmation of the null hypothesis in the case of H3. In the experiment, the notion of brevity was limited to number of disjuncts, omitting negation, and varying only between length 2 or 3. Longer or more complex descriptions might evince different tendencies. Nevertheless, the results show a strong impact of Coherence, compared to (a kind of) brevity, in strong support of the Local Conceptual Coherence Constraint.

6.9 Summary and outlook

This chapter has presented six experiments investigating a hypothesised constraint on plural reference. The starting point for the investigation was the notion of lexical similarity, and its relationship to the hypothesis that pluralities are represented as holistic discourse entities. The latter proposition is backed by previous psycholinguistic work. The other part of the hypothesis on Local Conceptual Coherence, dealing with similarity of categorisation of referents, was made more precise by testing several definitions of similarity. A distributional definition was found to be the best predictor of people's tendencies, possibly because it encompasses a variety of common elements in the semantic schema belonging to two words, where the schema reflects patterns of word usage in specific grammatical configurations. This take on the problem bears some relationship to work in Scenario Mapping theory by Sanford and Moxey (1995, 1999), which holds that pluralities are easier to comprehend when it is easy to conceive of some scenario in which they can be jointly mapped. While the experiments here did not directly investigate the manner in which these mechanisms operate, the results are at least partially compatible with a priming explanation, whereby the retrieval of lexical items primes the retrieval of related items. In the case of plurals, this may be one of the factors underlying the finding that similar or related categorisations of entities facilitate understanding and production.

The results of the experiments also afford a pragmatic interpretation that ties in with previous work by Kronfeld (1989) and Aloni (2002): the use of a plural reference by a speaker or author carries the implicature that there is some relevant link between the elements of the set. Under this perspective, therefore, production of a plural referring expression referring to a plurality which has insufficient 'conceptual glue' to hold it together violates the expectation of relevance on the part of a listener.

Finally, I proposed to view Local Conceptual Coherence as characterising a family of algorithms, and compared the LCC model with the dominant, Gricean model that has inspired most work on plural reference. The results of this experiment do not falsify the brevity-oriented model, but they do not offer support for it either. Specifically, there is no evidence that readers prefer brief descriptions compared to non-brief ones when the two descriptions are equally conceptually coherent, in the sense defined in this chapter. In contrast, local coherence was shown to exert a strong influence, and often tipped the balance in favour of the one description out of two possible alternatives that covered a set of referents in the most conceptually coherent way.

These experiments will serve as the foundations for the algorithmic work in the following chapter, where I propose two different algorithms that instantiate the LCC family. Porting the results of these experiments to the generation scenario raises a number of questions. First, supposing that the domain does not permit the satisfaction of the Conceptual Coherence constraint, should the algorithm terminate with failure? The reply to this question should of course be negative (among other things, there is no evidence that people fail when elements of a plurality cannot be similarly conceptualised). The second question has to do with the relationship between *properties* – semantic objects – and words, a relationship already alluded to at the beginning of this chapter, and one which has a bearing on the related problems of the generation gap (Meteer, 1991), and Logical Form Equivalence (Shieber, 1993). A lexically-driven strategy is supported by the results of Experiment 1, which found that WordNet-based similarity measures correlate less highly with people's judgements of likelihood of usage of disjunctive noun phrases. Other authors have called for the design of "lexically-aware" content-determination modules to avoid violation of noun-modifier combinatorial constraints (e.g. Lapata et al., 1999), which I suggested in §6.3 is another type of local constraint on noun phrase formulation.

Viewing the Local Conceptual Coherence Constraint at the lexical level (or rather, blurring the boundaries between the strictly lexical, and the purely semantic levels) is also compatible with the theoretical framework outlined in §6.2.4, for the reasons discussed there. However, lexical items and properties are difficult to talk about in the same terms. Do lexical items have an extension, or is it more correct to say that lexical items denote properties or concepts which have an extension? These, and related matters, form the central questions of Chapter 7.

Chapter 7

Generating conceptually coherent descriptions

7.1 Introduction

This chapter brings together the work from the previous two into an integrated framework which takes into account the kinds of local coherence and similarity constraints for which empirical evidence was found. Three such constraints will be addressed, the first of which is the primary goal that the algorithms described below seek to achieve:

1. Maximise similarity between categorisations of elements of a set of referents. This is interpreted as a constraint on noun-noun similarity in plural coordinate NPs. By hypothesis, the case where a plural NP does not involve disjunction (e.g. *the professors* vs. *the professor and the lecturer*) is a limiting case of this constraint, since all referents are identically categorised.

2. Since categorisations make some aspects of an entity more salient than others, further information predicated of an entity using (adjectival) modifiers should take this into account. This is interpreted as a constraint on maximising the collocational probability between the head noun of an NP, and the adjectives that are selected to further describe the set it denotes (cf. Cruse, 1986; Murphy, 1990; Lapata et al., 1999).

3. If same-head NPs are generated, they should only be coordinated if aggregating them will not violate the syntactic complexity limitations found in the empirical study in §5.5 (p. 154). Moreover, coordination of adjectives within an NP (*the red and blue chairs*) is constrained by similarity. This was interpreted in Chapter 5 as a constraint on coordinating values of the same attribute. Here, this is generalised to a constraint on adjective-adjective similarity, in line with the framework adopted.

This chapter will describe two algorithms that meet the requirements above. These should be considered as two possible algorithmic interpretations of the family of algorithms which take the Local Conceptual Coherence constraint (LCC) as their starting point.

One result that was regularly obtained in the Magnitude Estimation experiments of §6.5 (p. 178) was that a distributional definition of similarity, based on the occurrence of words in particular grammatical contexts, was the best predictor of people's preferences in comparison to other measures which were taxonomy-based, and to a certain extent, other notions of conceptual relatedness which relied on intuitive categorisations (for example, *human* vs. *artifact*). This measure

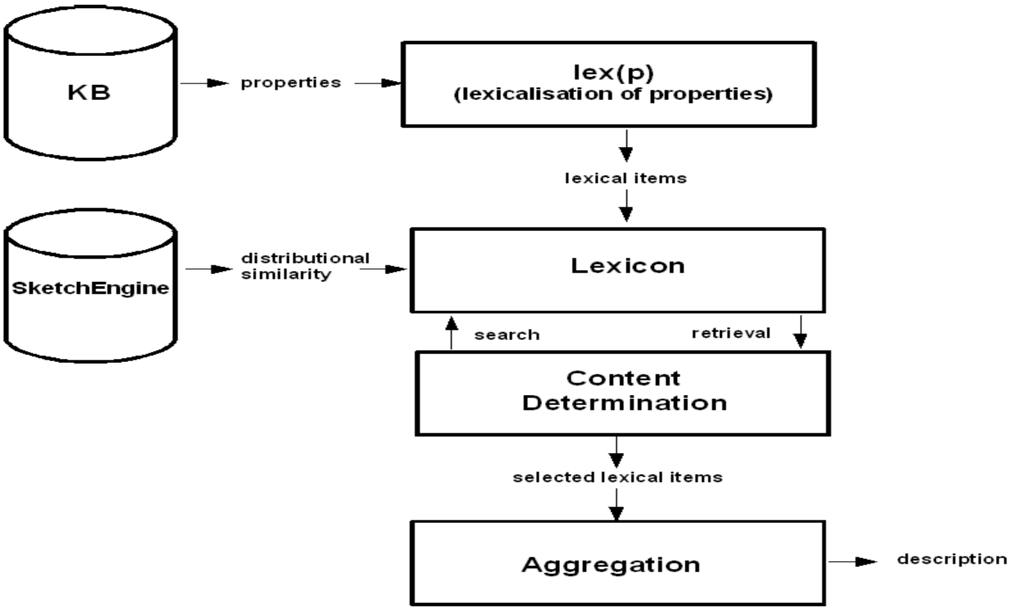


Figure 7.1: Basic architecture of the system

of similarity was then shown to be a reliable predictor of people's aggregation and content determination decisions. Therefore, the algorithms described below take this definition of similarity as their starting point. Moreover, because the most adequate definition of similarity found is *lexical*, based on the usage of words rather than concepts or senses, the content determination procedures described in this chapter are **lexically-driven**, in that the search space of these algorithms is no longer populated exclusively by properties, or attribute-value pairs, but by lexical items.

Lexically-driven generation does not constitute an abandonment of the extensional success criterion for GRE, which was the starting point for this thesis, and which relies heavily on the notion of denotation or extension of a property or formula. This criterion is still required to ensure that a distinguishing description is returned whenever one exists. Lexical items in the present framework are broadly conceived as property-to-word mappings, where 'properties' are the equivalent of 'concepts' underlying lexical items. Because of the existence of this mapping, it will still be possible to talk of the 'extension of a property'. However, there is now a new level, namely, the 'realisation of a property'. It is at this level that the notion of similarity acquires its importance, and it is also this level that will drive the description-building process. To immediately clarify the framework, Figure 7.1 displays the basic architecture to which the (implementations of) the algorithms described here conform:

1. The Knowledge Base is still assumed to be the repository of information about the universe of discourse.

2. Properties in the Knowledge Base are mapped to lexical items, which reside in the Lexicon. The mapping is obtained via a function $lex(p)$, described below.

3. The Lexicon is a structured repository of lexical information, backed by distributional similarity data obtained from the SketchEngine database (Kilgarriff, 2003). Using this information, it becomes possible to estimate the 'semantic neighbourhood' of a lexical item, that is, the relative semantic distance between one item and others. Pairwise distance between

lexical items is a function of their similarity, as defined in equation (6.8) (p. 178).

4. Content determination is now conceived as a search and retrieval process, which takes lexical items from the Lexicon, and constructs descriptions incrementally as such items are retrieved. Items which are retrieved and selected feed into an aggregation component. The form-meaning mapping between properties and words allows an algorithm to take both extensional and word-based aspects into account.

As discussed in Chapter 2 (see especially §2.7.7, p. 63), the consideration of aspects of realisation in GRE is hardly alien to the area – work on anaphora (Krahmer and Theune, 2002), gradable properties (van Deemter, 2006) and even, to some extent, plurals (Horacek, 2004) has in the past contained proposals motivated by similar broad concerns. The lexicalist framework adopted here is most closely related to that of Siddharthan and Copestake (2004), whose greedy algorithm attempts to minimise the ambiguity of a description in context by taking into account semantic relationships between a property selected for a referent and those of its distractors. Nevertheless, the aims of this procedure were quite different; it was not similarity that was the focus of the work of Siddharthan and Copestake, and ambiguity was operationalised using WordNet.

Another feature of the algorithms described here is that they maintain the basic Content Determination strategy outlined in Chapter 5. In describing an arbitrary set, the task is broken down into sub-tasks, using an opportunistic partitioning strategy. Once aggregation is included in the picture, this becomes a cycle of description-building and merging of new content with existing content.

The remainder of this chapter is structured as follows. The first part (§7.2.3, p. 212) begins with a description and formal, graph-theoretic definition of the Lexicon. The Lexicon is the fundamental data structure, in which lexical items are defined as mappings between KB properties (semantic objects) and words or strings (lexicalisations), while such items are connected by edges that reflect their similarity. For the purposes of the present chapter, attention is restricted to two kinds of lexical items, namely nouns and adjectives, each of which is organised as a subgraph of the Lexicon.

§7.3 (p. 214) then outlines the Content Determination process in skeletal form; this sketch forms the basis for the two algorithms to be described next. In this section, I discuss how the earlier partitioning algorithm of Chapter 5 is incorporated into the new model, and how the search space for Content Determination is populated by lexical items. Given that these items are property-word mappings, so that KB information (specifically, extensionality) is still available, the generalisation to the new framework is quite simple. Later in the section (§7.3.1, p. 217) I revisit the results reported in §5.5 (p. 154) on same-TYPE aggregation for premodifiers, discussing how the semantic and syntactic constraints found in that corpus-based study and incorporated into the aggregation procedure described there, can now be merged within the architecture of Figure 7.1.

Having thus set the stage, §7.4 (p. 218) is concerned with formulating a precise definition of Local Conceptual Coherence. In terms of the discussion of Chapter 2, from which all other discussions of GRE algorithms took off, this is the point where I define a preliminary ordering or 'adequacy' relation between alternative descriptions of a set of referents. Since the Lexicon is a weighted graph with edges reflecting the similarity between lexical items, there is a straightforward sense in which the Conceptual Coherence of a description depends on the semantic distance between the lexical items in the description. It turns out, however, that under this definition, finding the optimal description requires an exhaustive search, so that tractability issues again rear their heads. In a way that parallels the earlier discussion of computational interpretations of Gricean brevity (§2.5, p. 33), I then propose a weaker interpretation of the LCC (§7.5, p. 218), one which immediately points the way towards a greedy solution. Since the Lexicon is structured as a graph, the discussion of greedy algorithms takes as its starting point some well-known greedy solutions to approximate shortest connection networks in connected graphs. I offer a new conception of 'greed', one which views a description (a set of lexical items) as a subgraph of the Lexicon, and selects the next item to be tested for inclusion based on the overall semantic distance between items in the description. This model is applied to the selection of both nouns and adjectives.

At this point, two further incarnations of the greedy model are discussed. The first of these (§7.6, p. 226), is based on a pre-compilation step, in which a clustering algorithm is used to group together lexical items by their semantic relatedness. I call the resulting clusters *perspectives*; the sense in which this term is used here should be understood as reflecting entirely lexical forces. A perspective under this definition is a group of lexical items that, given the corpus-derived similarity information, are known to be usable in the same linguistic contexts. The second model to be discussed (§7.7, p. 230) is based on a different kind of solution. Rather than first finding perspectives and seeking to minimise the distance between perspectives reflected in the description, this model explains the LCC in terms of lexical priming. Whenever a lexical item is selected, it 'activates' its neighbours in semantic space, and the activation is a function of how semantically related those neighbours are to it. §7.8 (p. 233) discusses the different takes on the problem that these two models represent, also linking this discussion to some of the theoretical and empirical background introduced in Chapter 6. In particular, my focus on lexical forces ('local' constraints) raises the question of how these interact with 'global' constraints, that is, communicative intentions which play a causal role in how a speaker selects content for a description. I do not pretend to have a solution to this question, but I argue that any solution that incorporates global constraints also needs to take into account both the ground-level, bottom-up processes dealt with in this and the previous chapter.

7.2 Lexical items and the Lexicon

There are two types of lexical item, namely Noun ($\mathcal{N}$) and Adjective ($\mathcal{A}$). The basic structure of these is shown below, where $\mathbb{P}$ is the set of KB properties and $\mathcal{W}$ is the set of word-forms available.

(7.1)
$${}_{\textit{Noun}}\begin{bmatrix} \text{CAT} & \mathcal{N} \\ \text{SEM} & p \in \mathbb{P} \\ \text{LEMMA} & w \in \mathcal{W} \end{bmatrix}$$

(7.2)
$${}_{\textit{Adjecive}}\begin{bmatrix} \text{CAT} & \mathcal{A} \\ \text{SEM} & p \in \mathbb{P} \\ \text{LEMMA} & w \in \mathcal{W} \end{bmatrix}$$

Every lexical item has three fields. The first is a category CAT, which is $\mathcal{N}$ or $\mathcal{A}$. The second is a SEM field. Lexical items constitute a mapping between KB properties and words. Therefore, SEM is a KB literal, an element of the set $\mathbb{P}$ of properties in the KB. These objects are also specified as having a word-form LEMMA. The actual mapping of properties to lemmas is carried out by the lexicalisation function $lex(p)$ in Figure 7.1, which is defined as follows:

$$lex(p) : \mathbb{P} \rightarrow \mathcal{P}(\mathcal{W}) \tag{7.3}$$

For the work described here, the function $lex(p)$ was realised using WordNet. The KB contains properties whose values can be represented as WordNet senses for the purposes of lexicalisation. $lex(p)$ returns the set of possible realisations of a given WordNet sense, that is, the elements in its WordNet synset. This use of WordNet was primarily driven by necessity, since it is a sizable repository of near-synonyms.

Note that the property-to-word mapping in (7.3) is not a bijection, since a property can be mapped to several words. As an example, suppose the KB contains the property $\langle$OCCUPATION : *chemist*$\rangle$, and the relevant sense of this property is ‘pharmacist’ (sense #2 in WordNet). The function $lex(p)$ returns the realisations $\{$*chemist, druggist, pharmacist, apothecary, pill pusher, pill roller*$\}$. For each of these six realisations, there is a separate lexical item in the Lexicon, whose SEM is the property $\langle$OCCUPATION : *chemist*$\rangle$, and whose LEMMA is the realisation. Two such lexical items are exemplified below in (7.4) and (7.5).

(7.4)
$$ {}_{\textit{Noun}}\begin{bmatrix} \text{CAT} & \mathcal{N} \\ \text{LEMMA} & \text{chemist} \\ \text{SEM} & \Big\langle \text{OCCUPATION} : \textit{chemist} \Big\rangle \end{bmatrix} $$

(7.5)
$$ {}_{\textit{Noun}}\begin{bmatrix} \text{CAT} & \mathcal{N} \\ \text{LEMMA} & \text{pharmacist} \\ \text{SEM} & \Big\langle \text{OCCUPATION} : \textit{chemist} \Big\rangle \end{bmatrix} $$

The grammatical category of a lexical item is determined using a lookup table consisting of word-category pairings obtained from the BNC.[1] Distinguishing between the categories of lexical items is crucial for achieving Local Conceptual Coherence, since the three sub-goals of this constraint outlined at the beginning of this chapter are related to noun-noun, noun-adjective and adjective-adjective similarity or collocational relatedness. Therefore, the Lexicon is conceived as a structured repository in which nouns and adjectives are separately represented, though linked. More formally, the Lexicon is a directed, bipartite graph, which connects nouns to adjectives. Nouns form the nodes of a graph called a **Noun Graph**, abbreviated $L_{\mathcal{N}}$, while adjectives are held in an **Adjective Graph** $L_{\mathcal{A}}$.

Given the representation of the two types of lexical item in (7.2) and (7.1), I will often use the following notation in what follows:

CAT(l) is the category of lexical item l;

[1] Word forms in the lookup tables are lemmatised using the Sussex Morphological Analyser (Minnen et al., 2001).

$\textsc{sem}(l)$ is the semantic representation of l;

$\textsc{lemma}(l)$ is the word-form paired with $\textsc{sem}(l)$ in l;

$[\![\ \textsc{sem}(l) \ ]\!]$ is the extension of l, that is, the extension of the attribute-value pair that l denotes.

7.2.1 Nouns

The Noun Graph of the Lexicon contains information about the semantic similarity between nouns. Similarity estimates are obtained from the SketchEngine database, and are based on the definition of similarity by Lin (1998b,a). Grammatical triples for these calculations are obtained from the BNC. The similarity between two words is denoted $\sigma(w, w')$, as per the usage in the previous chapter. In the nominal component of the Lexicon, nouns are connected by edges whose weights represent the semantic distance between them, which is calculated by taking the multiplicative inverse of their pairwise similarity, normalised to deal with possible zero values. Just as similarity ranges in $(0, 1)$, where 1 indicates perfect similarity, distance also ranges in $(0, 1)$, where 1 is the maximal semantic distance that can hold between a pair of nouns. Edges in the Noun Graph are undirected, because the definition of similarity used is symmetric, so that it is sufficient to represent the distance between two nodes as the weight on a single edge (rather than two directed, weighted edges).

Definition 8. Noun Graph

A Noun Graph $L_{\mathcal{N}}$ is a connected, undirected, weighted graph $\langle N, E_{\mathcal{N}}, \delta_{\mathcal{N}} \rangle$ where:

- $N = \{l | \textsc{cat}(l) = \mathcal{N}\}$
- $E_{\mathcal{N}} \subseteq N \times N$
- $\forall \langle w, w' \rangle \in E_{\mathcal{N}} : \delta_{\mathcal{N}}(w, w') = \frac{1}{1+\sigma(w,w')}$

The Noun Graph is the main component of the lexicon as far as Content Determination is concerned. To select properties for a referring expression, the content determination algorithms discussed below attempt to minimise distance between those nodes of the graph selected for inclusion in the description.

7.2.2 Adjectives

Two kinds of lexical relationships involving adjectives are relevant. Adjective-adjective similarity is used to account for the constraint that coordinated adjectives within an NP tend to be similar. This was found to be the case in Chapter 5 (see §5.5, p. 154). Based on experimental and theoretical work by other authors, I have also proposed that noun-adjective relatedness is important since, once a referent has been categorised, the categorisation makes some aspects of that referent more salient than others.

To account for adjective-adjective similarities, adjectives too are represented in a connected, undirected, weighted graph, defined as a triple $\langle A, E_{\mathcal{A}}, \delta_{\mathcal{A}} \rangle$. Its definition is identical, mutatis mutandis, to that of the Noun Graph in Definition 8.

Noun-adjective relatedness is captured by the connection between the two graphs $L_{\mathcal{N}}$ and $L_{\mathcal{A}}$. Recall that similarity of two words as defined by Lin depends on prior estimates of the mutual information between each of the two words, and other words in specific grammatical contexts.

thin		long-haired	
noun	**salience**	**noun**	**salience**
layer	40.44	cat	16.67
section	32.36	king	16.17
strip	30.22	player	12.37
air	28.57	veteran	6.44
line	26.88	professor	5.5
ice	23.15	biker	8.16
moustache	21.63	blonde	7.74
man	21.24	student	3.32

Table 7.1: Noun-adjective collocational salience

	BASE TYPE	OCCUPATION	SPECIALISATION	GIRTH	HAIR-COLOUR	SENIORITY
e_1	man	professor	biologist	fat	dark-haired	senior
e_2	man	lecturer	geologist	chubby	blond	assistant
e_3	woman	lecturer	physicist	thin	blonde	senior
e_4	woman	pharmacist	–	fat	dark-haired	–
e_5	woman	doctor	psychiatrist	thin	dark-haired	consultant
e_6	man	lecturer	chemist	chubby	blond	assistant
e_7	woman	professor	biologist	thin	dark-haired	emeritus

Table 7.2: The input to lexicalisation

Given a word w and a grammatical relation rel, the mutual information $I(rel, w, w')$ gives an estimate of the *salience* of w' as a co-argument of w in rel. The part of the LCC that deals with dependencies between categorisation and modification is operationalised by taking into account, for each available noun in the lexicon, and each available adjective, the mutual information of the adjective and the noun in the grammatical relation of pre-modification. This gives an indication of how salient an adjective is with respect to a noun. In what follows, this will sometimes be abbreviated as $sal(a, n)$, which is to be read as 'the salience of adjective a with respect to noun n in the pre-modification relation'.

Table 7.1 shows some examples of the salience of the adjectives *thin* and *long-haired* in relation to some nouns. Words such as *student* and *professor* tend not to be modified too frequently by *long-haired*, in comparison to *player* and *biker*. Similarly, *thin* is a very salient premodifier of *man*, but less so of *professor*, among whose top premodifiers *thin* does not occur. The salience estimate of an adjective in relation to a noun makes most sense when considered as a relative estimate of the frequency with which the noun is modified by the adjective, in comparison to other adjectives. Thus, while *long-haired* is a more salient modifier of *biker* than it is of *professor*, the latter has, among its most salient premodifiers in the BNC, the words *emeritus*, *associate*, *visiting* and *retired*.

7.2.3 The lexicon data structure

Given the relationships that obtain between pairs of nouns and adjectives, a Lexicon can be formally defined.

Definition 9. Lexicon

A Lexicon $\mathcal{L}$ is a directed, weighted, bipartite graph $\langle L_{\mathcal{N}}, L_{\mathcal{A}}, E_{\mathcal{L}}, \delta_{\mathcal{L}} \rangle$ where:

- $L_{\mathcal{N}} = \langle N, E_{\mathcal{N}}, \delta_{\mathcal{N}} \rangle$ is a Noun Graph;
- $L_{\mathcal{A}} = \langle A, E_{\mathcal{A}}, \delta_{\mathcal{A}} \rangle$ is an Adjective Graph;
- $E_{\mathcal{L}} \subseteq N \times A$
- $\forall \langle n, a \rangle \in E_{\mathcal{L}} : \delta_{\mathcal{L}}(n, a) = \frac{1}{1+sal(a,n)}$

By this definition, the Lexicon represents a function taking a noun-adjective pair and returning a collocational salience value. To take an example of how a KB is realised in this data structure, consider Table 7.2, which represents a KB in $\langle$A : *v*$\rangle$ format. To simplify the presentation, it lists the values of the first three attributes as 'nominal' properties, that is, properties which will find their way into the Noun Graph of the Lexicon. The others are 'adjectival'. Assuming that $lex(p)$ is realised using WordNet, as explained above, this will yield a large number of possible lexical realisations of the various properties. Some example realisations are given below:

1. lex(*professor*) = $\{$*professor, prof*$\}$
2. lex(*lecturer*) = $\{$*lector, lecturer, reader*$\}$
3. lex(*pharmacist*) = $\{$ *pharmacist, druggist, chemist, apothecary, pill pusher*$\}$
4. lex(*doctor*) = $\{$*doctor, doc, physician*$\}$

Lexicalisations of a given property, at least in the manner carried out here, will result in words with various shades of meaning. Thus, *pill pusher* or *prof* are somewhat more informal terms than the corresponding entries in the KB. The idea of representing lexical information as a connected graph is to capture these shades of meaning through associations between lexical items. Another feature of the current example KB is that there is a value of OCCUPATION (*pharmacist*) which has, as one of its possible lexicalisations, the word *chemist*. However, there is also a value of SPECIALISATION (the value true of e_6) which will have *chemist* as one of its realisations. The two $\langle$A : *v*$\rangle$ pairs correspond to different 'senses' of the word: one is a practitioner, who works in a pharmacy (this corresponds to sense $\#2$ in WordNet); the other is an academic, research chemist (sense $\#1$). Therefore, lexical ambiguity can also arise in the current framework.

A partial representation of the Lexicon graph generated from this simple KB is displayed in Figure 7.2. This shows some of the noun-noun and noun-adjective links (adjectives are in rounded cells). For clarity, rather than the semantic distance defined in Definition 8, the edges of the graph in the figure show the actual similarity values between noun pairs. The similarity relationships between nouns can be a way of disambiguating their meanings or senses: *chemist* and *pharmacist* are both close, semantically, to *physician*. However, *chemist* is also very close to *physicist*, while the link between *physicist* and *pharmacist* is very weak.[2] By hypothesis, generating a description such as *the physicist and the chemist* is unlikely to result in perceived ambiguity between a 'pharmacist' and a 'research chemist' sense of *chemist*. The same can be said for a description like *the physician and the chemist*, which disambiguates the word in the other direction. The semantic neighbourhoods of words can therefore serve to disambiguate them. This rather Firthian

[2] Zero values on an edge $\langle w_1, w_2 \rangle$ in the graph represent cases where w_1 was not found in the first 500 words of the SketchEngine thesaurus entry for w_2.

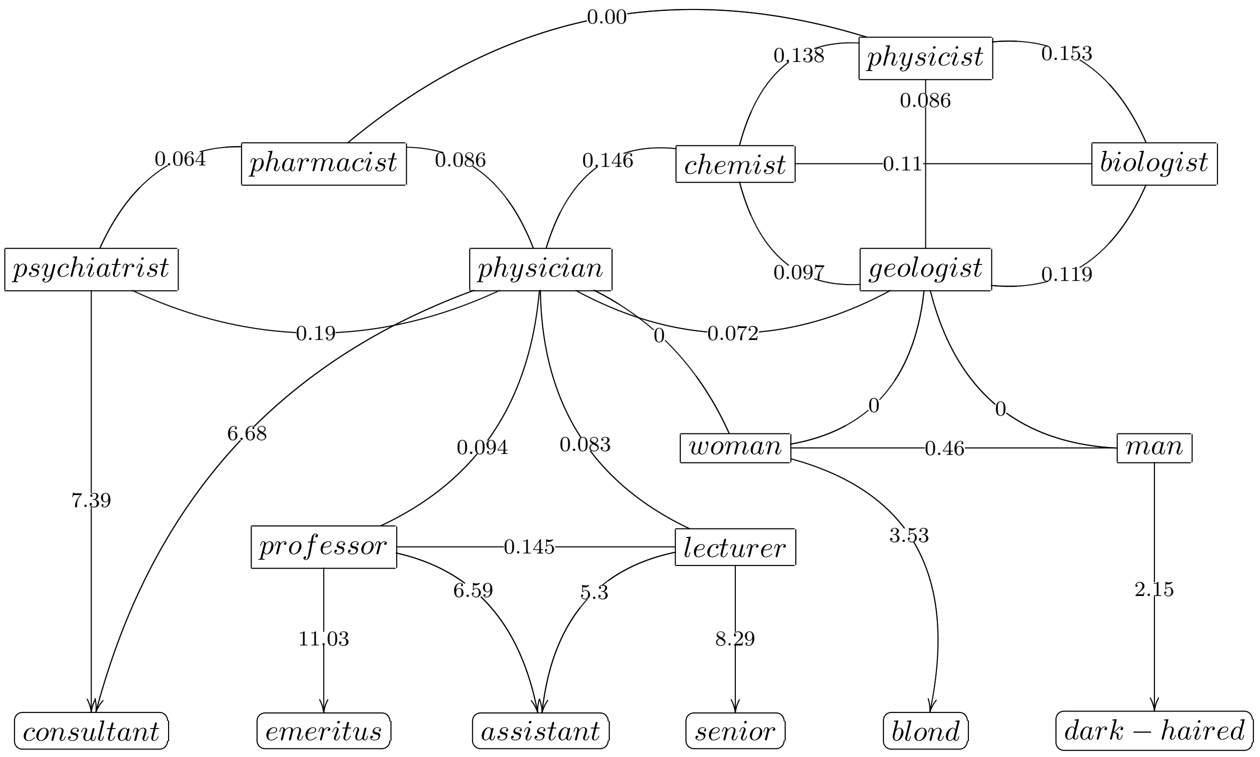


Figure 7.2: Graph-based representation of the Lexicon. Labelled edges indicate similarity values.

perspective on word meaning, which is also compatible with the view espoused by Wittgenstein in his later work (Wittgenstein, 2001), emphasises language use and is arguably a way of minimising ambiguity in generated descriptions. Just as, to use a well-worn example, the word *bank* would presumably not be perceived as ambiguous in the context of *the bank and the river*, so too, the use of *chemist* in conjunction with *doctor* (resp. *biologist*) may disambiguate it.

By explicitly representing these similarity relationships, the Lexicon indirectly represents the 'conceptual perspectives' available for the domain entities. These are more than simply attributes for which values are defined. As discussed in the previous chapter, a conceptual perspective, in the sense of the term used for example by E. Clark (1987; 1997a), carries with it a number of associative relationships, between a word and its context of use. These relationships are partially captured by the definition of word similarity used in the present work, where the similarity of words arises from 'talking about things in the same context'. Thus, the word *physician*, apart from being a value of OCCUPATION, bears strong associations with other words related to the medical profession, such as *psychiatrist*.

7.3 The generation process in outline

Having described the lexical component of the system, I will sketch how the generation process illustrated in Figure 7.1 functions. Generation is divided into the same two components as the partitioning algorithm of Chapter 5: a main procedure *makeReferringExpression* selects lexical items from the Lexicon, and updates the description via calls to the *updateDescription* procedure. The two procedures are now modified slightly to handle lexical items rather than properties. The process still makes use of Description Fragments (DFs), which represent a 'description-in-progress'. I use the generalised definition of DFs offered in §5.5 (Definition 7, p. 161) in what follows. To recap, DFs were defined as triples, consisting of a set of intended referents R_{DF}, a type property

T_{DF}, and a set of modifier properties M_{DF}. The latter was generalised to a set of sets of properties to enable the handling of aggregation of same-head NPs, where each set in M_{DF} corresponded to a disjunction of properties. In the current framework, this is altered slightly, so that T_{DF} is a noun, and M_{DF} an arbitrary set of (sets of) lexical items, which are either adjectives or nouns. Since a DF corresponds to an NP, pre-realisation, this is like stipulating that every noun phrase has a head noun, and can be pre- or post-modified by any number of adjectives or nouns. For instance, (7.6a) is a possible description of e_1 in Table 7.2; its counterpart in the present framework is (7.6b).

(7.6) (a) the senior professor who is a biologist

(b) $\left\langle \{e_1\}, \textit{professor}, \left\{ \{\textit{senior}\}, \{\textit{biologist}\} \right\} \right\rangle$

The partitioning algorithm introduced in Chapter 5 had two important characteristics. It used KB information to partition the set of referents R, and it maintained a description as a set of DFs, each of which represented an element of the partition. Aggregation was handled using Algorithm 4 (p. 162), which merged DFs with the same TYPE, modulo semantic and syntactic constraints.

In Chapter 5, similarity was interpreted as a constraint on using the same attributes for elements of a partition, as far as possible. This meant that redundant properties could be added to a DF. If a DF was complete, that is, it distinguished its intended referents R_{DF}, more information could still be added to it. In the current framework, the interpretation of similarity is directly related to lexical distributions in a corpus, so that the requirement is no longer to propagate attributes across DFs. Therefore, it is no longer necessary to keep adding information to a DF once it is complete. As a result, the algorithms described in this chapter make a distinction between those fragments which are still in progress, and those which have been completed. Incomplete fragments are kept in a data structure $Fragments$, while the description itself, D_{part}, consists of those fragments which, having been completed, are removed from $Fragments$ and added to D_{part}. Whenever this happens, the algorithm tries to merge a newly completed fragment to an existing one in D_{part}, using the aggregation algorithm. If this process fails, then the new fragment is simply added to the description. Thus, aggregation is carried out at the earliest possible stage.

The first step to making these issues more precise, is a definition of completeness for a DF.

Definition 10. Completeness of a Description Fragment

A DF $\langle R_{\text{DF}}, T_{\text{DF}}, M_{\text{DF}} \rangle$ is complete, abbreviated as $complete(\text{DF})$ iff:

- $T_{\text{DF}} \neq \perp$
- $[\![\text{ DF }]\!] = R_{\text{DF}}$.

No DF is complete unless it has a noun T_{DF} which maps to the head of the NP, and unless the DF distinguishes the subset of referents for which it is intended. Two further aspects of the content determination procedure introduced in Chapter 5 are retained here. The first is the notion of a 'distractor set', which was assumed to take the form of an array C, holding a set of distractors *for each element of* R. Thus, $C[r]$ for some $r \in R$ is the set of distractors of r given the current state of the content determination procedure. The second is the notion of contrastiveness, whose definition is reproduced below.

$$contrastive(p) \leftrightarrow \exists r \in R : \; C[r] - [\![\, p \,]\!] \neq \emptyset \tag{7.7}$$

Algorithm 5 Generation outline

```
Require: $R$   ▷ the intended referents
Require: $\mathcal{L}$   ▷ the Lexicon
 1: $D_{part} \leftarrow \emptyset$   ▷ initialise the description
 2: $Fragments \leftarrow \emptyset$   ▷ initialise the set of incomplete fragments

 3: procedure makeReferringExpression
 4:     while $(N \neq \emptyset) \vee (A \neq \emptyset)$ do   ▷ iterate through lexical items until they are exhausted
 5:         if $[\![ D_{part} ]\!] = R$ then   ▷ return as soon as referents are distinguished
 6:             return $D$
 7:         end if
 8:         $lex \leftarrow nextItem()$   ▷ retrieve the next lexical item
 9:         if $contrastive(\text{SEM}(lex))$ then   ▷ $lex$ has some contrastive value
10:             $R' \leftarrow R \cap [\![ \text{SEM}(lex) ]\!]$   ▷ initialise the set of referents included in $\text{SEM}(lex)$
11:             $updateDescription(R', lex)$   ▷ update the description, performing aggregation if possible
12:         end if
13:     end while
14:     return $D_{part}$   ▷ there are no more lexical items; return whatever has been generated
15: end procedure
```

A full discussion of how lexical items are selected by *makeReferringExpression* is the focus of the next section. However, it is useful to give a sketch of the main process at this point, to illustrate how descriptions are updated and maintained. Pseudocode for *makeReferringExpression* is shown in Algorithm 5.

The algorithm initialises two data structures at the outset. D_{part} is a description, initialised to $\emptyset$ [5.1], while $Fragments$ is the set of description fragments that are constructed as generation proceeds [5.2]. *makeReferringExpression* loops through lexical items until either both nouns and adjectives are exhausted [5.4] or the description D_{part} is distinguishing [5.5]. If neither conditions hold, then the algorithm selects the next lexical item via the function *nextItem* [5.8]. For the present, this will be glossed over; various kinds of selection heuristics for lexical items are discussed below. A call to *contrastive*$(\text{SEM}(lex))$ is made, whose argument is the *property* SEM associated with the lexical item [5.9]. If the lexical item has contrastive value, it is included in the description by a call to the update procedure.

Some further comments about the update procedure are in order. First, this procedure iterates through $Fragments$, performing much the same functions as before, namely to check whether a fragment is true of some elements of R', adding the lexical item to this fragment. In the original version, this procedure returned as soon as all referents in R' were accounted for (that is, the new item was added to all DFs which had a non-empty intersection with R'). In the algorithms discussed here, the procedure iterates through all elements of $Fragments$, performing one additional check:

```
if a DF is complete then
    remove DF from $Fragments$
    for each fragment DF$'$ in $D_{part}$ do
        if $aggregate(\text{DF}, \text{DF}')$ succeeds then
            remove DF$'$ from $D_{part}$
            update $D_{part}$ with the result of the aggregation
            break the loop
        end if
    end for
end if
```

In other words, any completed fragment is removed from the set of incomplete fragments, and an attempt is made to aggregate it with some other complete fragment, using the procedure *aggregate*(DF_1, DF_2) in Algorithm 4 (p. 162). If this succeeds, then D_{part} will contain the newly aggregated fragment. Otherwise, the newly completed fragment is added to D_{part}. Note that this doesn't affect the polynomial running time of the partitioning strategy. One result of only updating D_{part} with complete fragments is that there is no risk of returning a description which contains incomplete fragments in a domain where a complete description (defined as per Definition 10) exists. This is because the condition in [5.6] requires the description to refer to R: since fragments in D_{part} represent partitions, this condition will never be satisfied unless the referents are covered by a set of complete fragments.

7.3.1 Revisiting the semantic constraints on aggregation

The empirical study of plural NPs with coordinated adjectives, reported in §5.5.3 (p. 159) found that coordination of adjectives tended to be semantically constrained, so that the corpus contained NPs like *the red and blue chairs*, but not *the red and small chairs*. This was interpreted as a constraint on disjoining only values of the same attributes, within specific complexity constraints. The new framework, in which lexical items populate the search space of content determination, allows a more direct interpretation of the semantic constraints found in that section.

Coordinated adjectives in the BNC were highly similar on the distributional measure used in the current framework. This was independently backed by a positive correlation to a different similarity estimate, based on word glosses (Lesk, 1986). To make the decision of whether to aggregate two same-head descriptions, rather than check whether they represent values of the same attributes, the aggregation procedure in the new framework checks whether their pairwise similarity is sufficient to warrant coordination. The mean similarity reported for coordinate adjectives was .203. The adjectival component of the lexicon L_A represents adjective-adjective relations in terms of distance. Therefore, the new procedure only aggregates adjectives if their pairwise semantic distance does not exceed the following value:

$$\delta_A(a_1, a_2) = \frac{1}{1 + .203} \approx .8 \tag{7.8}$$

It is trivial to generalise the aggregation algorithm to deal with DFs containing lexical items, rather than properties. Using the above equation, the decision to aggregate – modulo syntactic complexity – becomes more straightforwardly derived from the results of the data analysis of (§5.5.3). By this method, given three description fragments such as (7.9) below, the algorithm will aggregate (7.9a,b) to yield *the senior and assistant professors*, but will not aggregate the third example, as *dark-haired* is very distant from both *senior* and *assistant*.

(7.9) (a) the senior professor

(b) the assistant professor

(c) the dark-haired professor

7.4 Preliminaries to Content Determination

To begin the discussion of content determination, I will first make the notion of conceptual coherence more precise, focusing first on conceptual categorisation, that is, on the nominals selected by an algorithm to describe a set. The following discussion harks back to the formal model of GRE of Chapter 2: the definition of conceptual coherence implicitly incorporates a definition of the ordering relation among alternative descriptions of a set of referents, with the best or maximally coherent description being the one that an algorithm should aim for.

Informally, a description of a set can be said to be **maximally coherent** if the semantic distance between the nominals in the description is minimal given the available lexical information. The first thing that needs to be defined is the notion of *conceptual distance* between elements of a description; based on this, a definition of **maximal conceptual coherence** becomes possible. Let D_{part} be a description (a set of complete description fragments) with $\mathbb{T}$ the set of nominals (T_{DF}) that head the DFs in D_{part}. The conceptual distance of D_{part}, abbreviated $dist_c(D_{part})$, is defined as follows.

$$dist_c(D_{part}) = \begin{cases} 0 \text{ if } |\mathbb{T}| = 1 \\ \sum_{\langle n,n' \rangle \in \mathbb{T} \times \mathbb{T}} \delta_N(n, n') \text{ otherwise} \end{cases} \tag{7.10}$$

By this definition, a description has zero distance if it contains only one complete DF, equivalent to the case where a description is singular, or morphologically plural without NP coordination. More generally, $dist_c$ is minimised the fewer disjuncts (fewer coordinates) a description has, so that the 'best' description is considered to be that which categorises referents in an identical fashion. In case of coordination, the distance is the sum of distances between the lexical items that head the coordinated NPs, here operationalised as the elements T_{DF} of the DFs that map onto the NPs.

Definition 11. Maximal Local Coherence (Strong version)
A partitioned description D_{part} is maximally coherent iff there is no description D'_{part} coextensive with D_{part} such that $dist_c(D_{part}) > dist_c(D'_{part})$.

In other words, to find a maximally coherent description, a content determination algorithm would have to ensure that the distance between conceptual categories represented in a description is absolutely minimal. To achieve this, the algorithm would have to search exhaustively through all possible combinations of nouns (and their associated modifiers) to find the best possible combination which will also distinguish the intended referents. Even if we ignore adjectival modifiers for the moment, this will mean finding the shortest connection network (sometimes called a Steiner network) in the Noun Graph, a known intractable problem (e.g. Cormen et al., 2003), because of the combinatorial explosion in the search space that it incurs.

7.5 Greedy approximations to Maximal Local Coherence

There are a number of well-known ways of getting around the intractability of shortest connection networks in fully connected, undirected graphs, many involving greedy algorithms. Applied to the current domain, a greedy solution to the Local Coherence problem would entail finding a description which approximated, though not necessarily strictly satisfied, Definition 11. I therefore begin by weakening this definition.

Definition 12. Weak Local Conceptual Coherence
A description D_{part} is weakly conceptually coherent iff there is no D'_{part} coextensive with D_{part}, obtained by replacing one noun in D_{part} with another noun in D'_{part} such that $dist_c(D_{part}) > dist_c(D'_{part})$.

Any description that satisfies Maximal Local Conceptual Coherence will also satisfy Weak Local Conceptual Coherence. However, the new definition makes the problem tractable, for now, in order to satisfy the weak coherence requirement, an algorithm need not compare all possible combinations of nouns to describe a set. Rather, it need only keep track of which nouns have been selected already, ensuring that at any stage of content determination, the next noun to be selected is the one which minimises the distance $dist_c$ in the description as it is so far. This approximation to the earlier, stronger definition of the LCC has a strong historical parallel to the original motivation for Dale's (1989) Greedy Algorithm, which approximated Full Brevity given that the latter was intractable (see §2.5, p. 33).

Among the best known greedy solutions to Shortest Connection Networks are a number of algorithms for constructing a Minimum Spanning Tree (MST) from a connected, weighted graph (e.g. Prim, 1957). A Spanning Tree of a graph G is defined as a sub-tree of G containing all the nodes in G. For weighted graphs, an MST is a spanning tree whose total weight is less than or equal to[3] that of any other spanning tree of G.

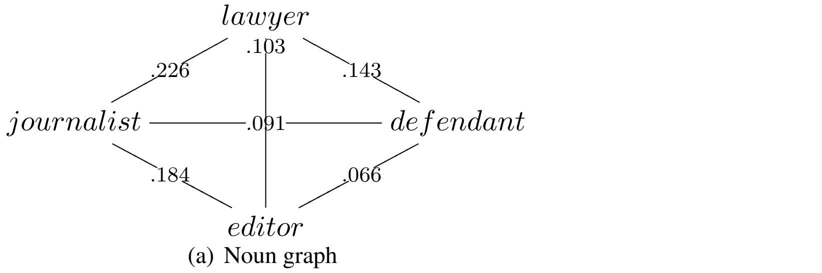
Prim (1957) proposed a polynomial-time MST algorithm. It starts from an arbitrary node in the graph which is the designated root of the tree, and maintains an ordering among the remaining nodes, in ascending order of their *cost*. Initially, all nodes except for the designated initial node $root$ have cost set to ∞. Starting from $root$, the algorithm proceeds by selecting a node, and updating the cost of every remaining node. Let n_i be the currently selected node at iteration i of the algorithm. Update of the costs of each remaining node n proceeds as follows:

- if the the distance $\delta(n_i, n)$ in the graph is less than the current cost of n, then:
    - set the cost of n to $\delta(n_i, n)$;
    - set n_i as the parent of n in the tree

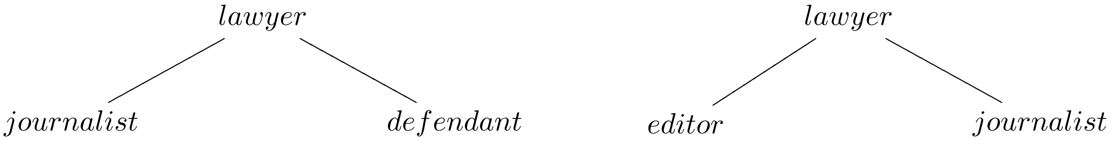
Consider what an application of Prim's strategy (or a similar one) to the Content Determination problem would entail. Rather than constructing a complete MST, the adapted algorithm would (a) select only those nodes of the graph which were useful because they excluded some distractors for some elements of R; (b) break the procedure and return as soon as a description was found to be distinguishing. However, the idea of only comparing nodes *locally* would be maintained.

There is first of all the problem of selecting the initial node (i.e. the initial Lexical Item in the Noun Graph). Suppose that the initial node lex_{root} is the one with the highest discriminatory power, that is, the lexical item whose associated SEM field was found to remove the most distractors. Focusing only on Nouns for the present, let $Nodes$ be the set of nodes selected by the

[3] Several MSTs may exist for a single connected graph.



(a) Noun graph



(b) Distinguishing subtree 1. Tree distance = 1.69

(c) Distinguishing subtree 2. Tree distance = 1.72

Figure 7.3: A violation of the triangle inequality. Edges are labelled with similarity values.

algorithm up to iteration i. This set and the remaining nodes in the Noun Graph represent a cut, that is, a partition of the set of nodes. Initially, $Nodes$ will only contain lex_{root}. At any point in the iteration, the next node to be selected is always guaranteed to be one which is semantically closest to (least distant from) *at least one* element of $Nodes$. For example, suppose the algorithm traverses the nodes of the Noun Graph in Figure 7.2, and assume that the intended referents are $R = \{e_1, e_2\}$ in Table 7.2. The algorithm might begin with the Lexical Item corresponding to *professor*, as this is true of e_1 and only includes the distractor e_7. The cost of every other lexical item is initially set to ∞. Once *professor* is selected, the lexical items are updated to reflect this, so that their costs reflect their distance from *professor*. The least distant item to be considered next would be *lecturer*. The description *the professor and the lecturer* is not distinguishing, and would require the addition of modifiers. The point of this partial example, however, is to illustrate how a greedy MST algorithm might be applied to the domains under discussion.

In general, a procedure to construct MSTs is not sufficient for the problem at hand. This is because the procedure selects nodes based on their distance or similarity to *one* nearest neighbour in $Nodes$. Consider a more complex case involving, for example, three intended referents, all of which are categorised differently (i.e. in three DFs). Call the nouns used to describe them n_1, n_2, and n_3. In order for a Prim-like algorithm to guarantee a (weakly) coherent description, the semantic distance between nodes of the Noun Graph would have to satisfy the following version of the triangle inequality:

$$\delta_N(n_1, n_3) \leq \delta_N(n_1, n_2) + \delta_N(n_2, n_3) \tag{7.11}$$

If this inequality didn't hold, the algorithm would risk selecting a set of nodes in which, for example, n_1 was the most highly related lexical item to n_2, and n_2 to n_3 in the Lexicon, but n_1 was semantically very distant from n_3. This is only a problem if there exists a better alternative whose overall cost was lower, that is, had a higher overall pairwise similarity between nouns. In general, the definition of similarity used here does not guarantee (7.11). Examples are not too difficult to come by. For instance, suppose the following KB represented part of a scenario involving three individuals in a court case. One of these, the defendant, happens to be the editor of a newspaper.

(7.12) (a) $\langle$PROFESSION : *editor*$\rangle$ (= e_1)

(b) $\langle$PROFESSION : *journalist*$\rangle$ (= e_2)

(c) $\langle$PROFESSION : *lawyer*$\rangle$ (= e_3)

(c) $\langle$ROLE : *defendant*$\rangle$ (=e_1)

The corresponding Noun Graph fragment is shown in Figure 7.3(a) (once again, labelled edges in this figure display similarities between noun pairs, rather than distance). The graph has two sub-graphs whose nodes would constitute distinguishing descriptions, namely $\{$*lawyer, editor, journalist*$\}$ and $\{$*lawyer, defendant, journalist*$\}$. The former has lower conceptual distance than the latter.[4] Figures 7.3(b) and 7.3(c) display two possible distinguishing sub*trees* of this graph, which might be selected by an MST-based algorithm. One tree containing all the nodes of the more costly subgraph turns out to be cheaper than that corresponding to the cheapest subgraph, hence would be the one preferred by an MST-based procedure.

All other things being equal, the Local Conceptual Coherence constraint would predict a preference for $\{$ *lawyer, editor, journalist* $\}$, trading off on the higher similarity between $\langle$*lawyer*, *defendant*$\rangle$, compared to $\langle$*lawyer*, *editor*$\rangle$, in favour of a lower overall cost. Put somewhat more generally, a greedy algorithm to approximate Maximal Local Coherence would have to consider a description as a subgraph, rather than a subtree, of the Noun Graph.

In addition to the inadequacy of trees as the target data structure for a greedy algorithm in this context, this example also illustrates another potential pitfall. A lot depends on the starting point for the algorithm. Thus, if the algorithm began from *defendant*, it would produce a description corresponding to *the lawyer, the defendant and the journalist*. In the present example, in fact, there is nothing to stop the algorithm from doing just this, if all the properties listed in the mini-KB in (7.12) are equally discriminatory. I will turn to this problem in the next section, dealing first with an extension of the greedy heuristic to sub-graphs rather than sub-trees.

Here is a variation on the same heuristic. We still assume that the algorithm starts from a designated root, and maintains a set $Nodes$ of lexical items which have been selected up to iteration i. This set represents the set of nouns in the description, on the basis of which the conceptual distance $dist_c$ is defined by (7.10). Thus, the algorithm maintains a cut of the graph $\mathcal{L}_N$. Rather than update the cost of remaining nodes by comparing them to their distance from the currently selected node, the algorithm always selects the node that will result in the least increase in the *overall* conceptual distance $dist_c$ of the description. Therefore, at any iteration through nominal nodes, the next item $next(L_{\mathcal{N}})$ is defined as follows:

$$next(L_{\mathcal{N}}) = \begin{cases} \arg\max_{n \in N} disc(n) \text{ if } Nodes = \emptyset \\ \arg\min_{n \in N - Nodes} \sum_{n' \in Nodes} \delta_N(n', n) \text{ otherwise} \end{cases} \quad (7.13)$$

where $disc(n)$ is the discriminatory power of lexical item n. The next node from $L_{\mathcal{N}}$ to be retrieved is the one that minimises the sum of distances between it and every other node in $Nodes$, unless $Nodes$ is empty, in which case, $next(L_{\mathcal{N}})$ returns the node with the highest discriminatory power, which functions as the root node for the search.

[4] This can be seen from the following calculation. The conceptual distance of $\{$*lawyer, editor, journalist*$\}$ is obtained by $\frac{1}{1+.103} + \frac{1}{1+.226} + \frac{1}{1+.184} \approx 2.56$. The same calculation for $\{$*lawyer, journalist, defendant*$\}$ yields ≈ 2.64.

One of the virtues of the greedy procedure discussed here is its efficiency. Rather than conduct exhaustive search and having to test all combinations of nominals in the worst case, it is guaranteed to find a set of nouns (or terminate with failure) in polynomial time. Let $n_{\mathcal{L}}$ be the number of nouns in the Noun Graph. This is also the maximum number of times the algorithm needs to iterate. Every time a node is added to the description, the algorithm will have to update the costs of the remaining nodes, in order to ensure that the next one selected satisfies (7.13). Suppose the algorithm selected n_d nodes. Then there are at most $n_{\mathcal{L}}$ updates that the main procedure $makeReferringExpression$ has to perform n_d times, which gives the algorithm complexity $O(n_{\mathcal{L}}^2 n_d)$.

This kind of greed is a partial solution to the problem noted in relation to Figure 7.3. The best (most coherent) distinguishing subgraph of the Noun Graph, with nodes $\{$*lawyer, editor, journalist*$\}$ would be found *if the algorithm began at the right node*. It would still not resolve the problem in the example if the root node of the search were *defendant*. This issue arises because of the two quite distinct goals that the algorithms under discussion are aiming to satisfy. On the one hand, finding contrastive properties is part of the basic problem definition; on the other, we require these properties to constitute a conceptually coherent cover of a set. Starting from a root node which has high discriminatory power targets the first goal, and is even a way to maintain brevity in a description; however, it may conflict with the second goal.

7.5.1 Dealing with modification

So far, I have focused on the kind of local coherence to do with nominal categorisation. However, nouns alone might not be sufficient to generate a distinguishing description. To deal with adjectival modifiers, the algorithm needs to be modified slightly, as follows. Recall that the selection of a new noun results in the projection of a noun phrase. The aggregation/realisation procedure discussed earlier will attempt to merge this with an existing phrase in a description fragment. If this does not account for all the referents that the phrase refers to, then it results in the construction of a new DF. Whenever a DF is found to be complete, it is merged with the final description. In case a new fragment is incomplete, modifiers can be added to it. These modifiers can be adjectives (*the fat professor*) or nouns in the equivalent of a relative clause structure (*the professor who is a biologist*). One way to extend the greedy algorithm to deal with adjectival modifiers is the following:

1. Search greedily through the set of nodes in the noun graph, until either all nouns are exhausted, or every element of R has been categorised, that is, there is a DF with a TYPE T_{DF} for all the referents;

2. At the end of this process, if there are incomplete fragments, begin a new iteration:

   (a) For each DF in $Fragments$, iterate through the adjectives in the Lexicon in descending order of their collocational salience with the head noun T_{DF}, adding an adjective if it is true of the referents referred to by the DF, and excludes some distractors.

   (b) If the adjectives are exhausted, iterate through the remaining nouns, adding a noun to the DF if it excludes some distractors.

This procedure will return a description whose components satisfy the second kind of local coherence that is of interest. Given a DF whose head noun is T_{DF}, which is not complete (i.e.

non-distinguishing for the subset of R that the phrase refers to), it will visit each adjective in $L_{\mathcal{A}}$ until the set is exhausted or the fragment is distinguishing. Let $Adj(\text{DF})$ be the modifiers included in a DF. At each stage of the iteration, the next adjective to be retrieved from the Adjective Graph is the one that satisfies the following condition:

$$next(L_{\mathcal{A}}, T_{\text{DF}}) = \underset{a \in A - Adj(\text{DF})}{\arg\min} \; \delta_{\mathcal{L}}(a, T_{\text{DF}}) \tag{7.14}$$

that is, the adjective in the Adjective Graph which is closest to the head noun of the fragment under consideration, and which is not already included in the fragment. This approach is somewhat wasteful, because it potentially requires the traversal of the set A, the nodes of the Adjective Graph, more than once. Nevertheless, it highlights a basic concern with prioritising nouns before other lexical items, because these are the fundamental building blocks of NPs. The same concern in Chapter 5 was easily dealt with, under the assumption that every referent had at most one value of TYPE. This assumption no longer holds, however.

One solution to this problem is to make lexical retrieval selective, and dependent on the state of the description being constructed. During the generation process, any fragment which is incomplete is given highest priority. The category of the next lexical item retrieved will depend on what the fragment with the highest priority requires. This makes $Fragments$, the set of DFs under construction, resemble a *chart* in chart generation systems (e.g. Kay, 1996). Such data structures are used to hold (normally syntactic) material under construction by a generator. New information is added to the chart on the basis of a priority function incorporated in an *agenda*, and combined to items on the chart on an opportunistic basis. The proposal made here has some resemblance to this procedure, and is also related to a proposal by Varges (2005a), which suggests treating a description in GRE as a chart of fragments, in an architecture that interleaves content determination with realisation. My approach, however, is not a full-blown chart-based solution. Rather, the idea is to maintain an ordering among fragments under construction, and retrieve from the Lexicon the item that is most likely to satisfy the requirements of the highest-ordered fragment.

The first ingredient towards achieving this is the notion of priority of a description fragment $\langle R_{\text{DF}} T_{\text{DF}}, M_{\text{DF}} \rangle$, which is defined straightforwardly as follows:

$$priority(\text{DF}) = \begin{cases} 0 \text{ if } complete(\text{DF}) \\ 1 \text{ if } T_{\text{DF}} \neq \bot \\ 2 \text{ otherwise} \end{cases} \tag{7.15}$$

Fragments which are complete have 0 priority, while those which lack a noun (that is, are not 'headed NPs') have highest priority, because the phrases they represent lack an essential element. To incorporate this new feature with $makeReferringExpression()$, a function is required that returns the maximum priority of a fragment. If this value is 2, then a noun is retrieved, otherwise, an adjective is retrieved. The revised procedure is shown in Algorithm 6. This time, the function $nextItem()$ is explicitly defined on the basis of the preceding discussion, and it makes use of the priority of fragments to determine whether a noun or an adjective is to be retrieved next.

The new procedure initialises the set $Nodes$ as well as the sets D_{part} and $Fragments$ [6.1–6.3]. Content determination proceeds by calling $nextItem()$ to return the next lexical item lex

Algorithm 6 Query and retrieval procedure

```
Require: $R$    ▷ the intended referents
Require: $\mathcal{L}$    ▷ the Lexicon
$D_{part} \leftarrow \emptyset$    ▷ initialise the description
$Fragments \leftarrow \emptyset$    ▷ initialise the set of incomplete fragments
$Nodes \leftarrow \emptyset$    ▷ initialise the set of nominal nodes

procedure makeReferringExpression
    while $(N \neq \emptyset) \vee (A \neq \emptyset)$ do    ▷ iterate through lexical items until they are exhausted
        if $[\![ \text{SEM}(D_{part}) ]\!] = R$ then    ▷ return as soon as referents are distinguished
            return $D_{part}$
        else
            $lex \leftarrow nextItem()$    ▷ retrieve the next lexical item
            if $contrastive(\text{SEM}(lex))$ then    ▷ $lex$ has some contrastive value
                $R' \leftarrow R \cap [\![ \text{SEM}(lex) ]\!]$    ▷ initialise the set of referents included in $\text{SEM}(lex)$
                $updateDescription(R', lex)$    ▷ update the description, performing aggregation if possible
                if $\text{CAT}(lex) = \mathcal{N}$ then    ▷ if $lex$ is a noun, insert it into $Nodes$
                    $Nodes \leftarrow Nodes \cup \{lex\}$
                    $N \leftarrow N - \{lex\}$    ▷ remove the noun from $N$; maintain a cut of the Noun Graph
                else if $\text{CAT}(lex) = \mathcal{A}$ then    ▷ if $lex$ is an adjective remove it from $A$
                    $A \leftarrow A - \{lex\}$
                end if
            end if
        end if
    end while
    return $D_{part}$    ▷ there are no more lexical items; return whatever has been generated
end procedure

procedure nextItem()
    $\langle R_{\text{DF}}, T_{\text{DF}}, M_{\text{DF}} \rangle \leftarrow \max_{\text{DF} \in Fragments} priority(f)$    ▷ retrieve the DF with the highest priority
    if $(priority(\langle R_{\text{DF}}, T_{\text{DF}}, M_{\text{DF}} \rangle) = 2) \wedge (N \neq \emptyset)$ then
        return $next(N_{\mathcal{L}})$    ▷ if this DF lacks a noun, return one if available
    else if $A \neq \emptyset$ then
        return $next(A_{\mathcal{L}}, T_{\text{DF}})$    ▷ otherwise return an adjective if available
    else
        return $\perp$    ▷ return null if all items have been visited
    end if
end procedure
```

[6.9]. This occurs within the body of a loop that terminates if there are no more lexical items, that is, the set of nominal nodes N and adjective nodes A have been exhausted [6.5].

The $nextItem()$ procedure begins by retrieving the fragment with the highest priority [6.25]. If this item has priority 2, then a noun is required. The return value of $nextItem$ is therefore the noun $next(L_{\mathcal{N}})$ [6.27] *unless* the set of nominal nodes N has been exhausted. Otherwise, if the set A of adjectives has not been exhausted, an adjective is returned [6.27], which is defined as the one with the highest salience (lowest distance from) the noun T_{DF} of the highest priority fragment. In case no further lexical items are available, it returns null [6.31]. In general, this will not happen because the main loop of $makeReferringExpression()$ breaks as soon as both sets of lexical items are empty. Whenever a lexical item is retrieved, it is removed from the node set of the corresponding graph [6.15–6.17]. Moreover, if the retrieved lexical item is a Noun, it is added to $Nodes$ [6.14]. This ensures that a cut of the noun graph is maintained, so that $Nodes$ and the set N of nominal nodes represent a partition.

Adjectives are also removed from A when retrieved. This means that any lexical item is retrieved at most once. Since the update of the description adds an item to any fragment to which it applies, this will never result in an item being ‘missed’. However, this procedure has an important consequence with respect to adjectives. An adjective that is retrieved has high collocational salience to the head noun of the fragment with the highest priority, but $updateDescription$ may also add it to another fragment. This weakens the requirement that adjectives have the highest possible collocational salience with respect to the nouns they modify, since adding an adjective to a low-priority DF does not guarantee that it has maximal salience in relation to its head. This problem is offset by the greedy maximisation of similarity between nominal heads: highly similar nouns will tend to share adjectives with high collocational salience, because premodifier salience is part of the definition whereby noun-noun similarity is estimated.

7.5.2 Worked example

To take an example of how this content determination process works, suppose we require a reference to $R = \{e_1, e_3\}$ in Table 7.2. The algorithm will begin by selecting the noun with the highest discriminatory value. In this case, this is *physicist*, which is entirely distinguishing for e_3. Updating the description results in the insertion into $Fragments$ of a DF, with $R_{\text{DF}} = \{e_3\}$ and $T_{\text{DF}} =$ *physicist*. $Nodes$ now contains *physicist*. On the next iteration, the most similar lexical item to *physicist* is retrieved (since this is the only element in $Nodes$). From the available items which have some discriminatory value for e_1, *biologist* is selected. The call to $updateDescription$ now has two outcomes. First, the only DF added to $Fragments$ so far is found to be complete. Thus, it is removed and merged with the description. D_{part} now consists of a single DF, as follows:

(7.16) $\langle\{e_3\}, \text{physicist}, \emptyset\rangle$

The set of referents is also updated, to reflect the fact that e_3 need not be described further. Second, a new DF is constructed, containing $R_{\text{DF}} = \{e_1\}$ and $T_{\text{DF}} =$ *biologist*. Hence, at the next iteration, an adjective is required because this DF has priority 2. The one with the shortest distance from *biologist* is *senior*. This completes the iteration, because *senior biologist* will be found to be complete at the next call of $updateDescription$. The outcome is *the physicist and the senior* *biologist*.

7.5.3 Interim summary

The greedy algorithm presented above is a tractable way to approximate maximal coherence. As described, it prioritises lexical items by category, based on the principle that reference requires categorisation first and foremost. This is operationalised in terms of the priority of a fragment and combined with (a) the opportunistic partitioning strategy incorporated in the update procedure, and (b) the aggregation strategy described in §7.3.

The heuristic whereby the root node for search is always the most discriminatory noun is one way of decreasing computational overhead, because it often means that entities can be distinguished earlier. However, it can have the undesirable outcome observed in relation to Figure 7.3. If the most discriminatory noun happens to be relatively distant from most other nouns which are required to distinguish the referents, then the outcome can be less than optimal. Another possible objection to the way generation is carried out is that, in treating nouns and adjectives differently, respecting the dependency that adjectives have on nouns, it is incurring increased processing. I have motivated this strategy on the basis of empirical and theoretical work that has shown how noun-adjective collocations affect the comprehensibility of a description. However, a separate account, one which focuses more specifically on reducing computational overhead as much as possible, is conceivable. This would hold that as long as the primary goal outlined at the beginning of this chapter – that of conceptual coherence in the way entities are (nominally) categorised – is satisfied, then the constraint on collocational probabilities can be relaxed. Thus, rather than prioritising fragments and selecting content based on their requirements, resulting in a comparison not only of nouns, but also of adjectives every time one of these is required, an alternative way of going about it would be to use the Lexicon graph to first identify clusters of items that constitute a conceptual perspective, then assign to each cluster the adjectives that go with the nouns in that perspective. This would have another desirable outcome, namely that it brings the computational strategy closer in spirit to the notion of perspective that psycholinguists have sometimes used to explain conceptual coherence (Clark, 1987, 1997a, 1991).

The next two sections will develop the greedy algorithm in two separate directions, both of which also represent ways of making it increasingly unlikely that the greedy search will be led astray by considerations of discriminatory power. The first is the one just mentioned: I describe a clustering algorithm and a slightly modified version of the content determination procedure. The net effect of clustering is twofold: First, there is the theoretically desirable outcome of finding the available perspectives in the Lexicon; second, there is also the potential outcome of reducing computation because clustering is effectively a way of reducing information.

An alternative approach is also proposed, whereby the Lexicon is viewed, not as a passive repository of lexical information (with relationships between its elements), but as an active repository in which items are activated to different degrees. This kind of model would view conceptual perspective as an emergent property of a more fundamental, low-level lexical priming mechanism.

7.6 Greedy search with information reduction

The basic idea behind information reduction is to find a way of structuring the knowledge in the Lexicon, grouping together related items into clusters that intuitively correspond to the available conceptual perspectives in a particular domain. Thus, the nouns in the graph in Figure 7.2 might be grouped into the following sets:

(7.17) (a) $\{$*professor, lecturer*$\}$
(b) $\{$*physician, pharmacist, psychiatrist*$\}$
(c) $\{$*chemist, physicist, biologist, geologist*$\}$
(d) $\{$ *man, woman* $\}$

I refer to these sets as 'conceptual perspectives' in that they group together items which shed light on a particular mode of conceptualising entities and, moreover, group together the available modes of categorisation which are compatible (that is, highly similar). If this can be carried out, then the information available in the lexicon is potentially reduced considerably. Once nouns are categorised in this way, adjectives can be added: each adjective in the Lexicon is assigned to a perspective which contains those nouns with which it is most highly related in terms of its salience value. Clusters can be related to each other in the same way that lexical items can, in the sense that some clusters are closer to certain others.

To the extent that the intuition behind the clustering in (7.17) is correct, it affords a generalisation. In each of the four clusters, words have been grouped with their nearest neighbours in the semantic space represented by the Noun Graph. Thus, the nearest neighbour of *professor* is *lecturer*, while *biologist* clusters with *physicist* and *geologist*, both of which are closer to it than to any other node. One possible issue arises with *chemist* which, due to the lexical ambiguity noted in §7.2.3, has *physician* as its nearest neighbour in the graph. It is, however, the nearest neighbour of *physicist*.

The Lexicon therefore provides the basic ingredients for a clustering procedure based on semantic distance, because the Noun Graph represents a semantic space $\mathbb{S} = \langle N, \delta \rangle$, where N is a set of lexical items (the nodes of the Noun Graph), and δ a distance function. The definition of a perspective makes use of the geometric notion of convexity (cf., e.g. Preparata and Shamos, 1985).

Definition 13. Perspective
A perspective $\mathcal{P}$ is a convex subset of $\mathbb{S}$, i.e.:
$\forall n, n', n'' : ((n, n' \in \mathcal{P} \wedge \delta(n, n'') \leq \delta(n, n')) \rightarrow n'' \in \mathcal{P})$

By this definition, if two lexical items are in the same perspective or conceptual cluster, and a third lexical item falls between them in terms of distance, then it too must be in the same cluster. All the clusters in (7.17) are convex in this sense. This is also true of the cluster containing the problematic *chemist*. As a glance at Figure 7.2 will show, there is no lexical item that falls between *chemist* and either of the nouns in its cluster which is not also in that cluster. To perform clustering, the algorithm will rely on the nearest neighbour of each lexical item in the Noun Graph, that is, the one which is closest to it. Let $nn(l, l')$ abbreviate 'l is the nearest neighbour of l''. The procedure used for information reduction is an algorithm described in Gatt (2006b), which groups lexical items by taking the transitive closure of the nearest neighbour relation.[5] Its decision on whether to

[5] The clustering algorithm was originally proposed to deal both with lexical semantic clustering and with spatial clustering. See Gatt (2006a) for a description of how this algorithm is used as the basis for the generation of spatial references.

include a lexical item in a cluster C is based on the following Nearest Neighbour Principle:

$$l' \in C \wedge nn(l, l') \rightarrow l \in C \tag{7.18}$$

Clearly, any procedure that follows (7.18) will yield convex subsets of the semantic space. Note that the nearest neighbour relation is non-symmetric. (For example, the nearest neighbour of *chemist* in Figure 7.2 is *physician*, but the latter has *psychiatrist* as its nearest neighbour.) Cases of symmetry, referred to as *reciprocal pairs* (Preparata and Shamos, 1985) can however arise, as witness *professor* and *lecturer*, which are mutual nearest neighbours. A further characteristic of the clustering algorithm is that it is not called upon to find a predefined number of clusters; rather, it is based on the assumption that there is some set $\mathcal{C}$ of clusters which are to be 'discovered'. This distinguishes it from some standard clustering algorithms, such as $k-$means, which require a preset number. As shown in Gatt (2006b), the output of the clustering procedure when applied to words closely approximates the groups produced by native English speakers when given the same words and asked to cluster them by their relatedness.

7.6.1 Content determination with conceptual perspectives

The clustering procedure described above is used to reduce the information in the Noun Graph, generating a new graph whose nodes are the available 'perspectives' in the domain. An example of the graph generated with the perspectives in (7.17) is shown in Figure 7.4. In the new perspective graph, the edges between nodes are weighted by the conceptual distance *between perspectives*. For any pair of perspectives $\langle \mathcal{P}, \mathcal{P}' \rangle$, the distance is obtained by the following formula:

$$\delta(\mathcal{P}, \mathcal{P}') = \frac{1}{1 + \frac{\sum_{\langle n, n' \rangle \in \mathcal{P} \times \mathcal{P}'} \sigma(n, n')}{|\mathcal{P} \times \mathcal{P}'|}} \tag{7.19}$$

that is, the mean distance between noun pairs across the two perspectives. As the Figure also shows, once clustering of the available categorisations of domain entities is carried out – a process involving the nominal part of the Lexicon – adjectives are added to each perspective. Each adjective is added to the perspective in the Lexicon which contains those nouns with which the adjective has the highest mean collocational salience. Note that the distance between clusters is often quite high, without large variations; this suggests that clustering by convexity yields an 'even' partition of the lexicon, so that items within a cluster are very close, and items across clusters are usually quite distant. However, distance between clusters would be more variable in domains in which the variety of lexical items was greater. For example, domains in which entities belonged to different ontological categories (e.g. *man* and *woman* versus *table* and *chair*) would yield greater distance between clusters.

Content determination now proceeds using the same greedy algorithm as before, but this time, the algorithm visits nodes that are themselves clusters of lexical items. For this reason, the original definition (7.13) of the *next* node to be visited by the algorithm at iteration i needs to be redefined. The new version of the algorithm still maintains a set $Nodes$, containing pointers to the perspectives visited at any stage of the iteration, from which the algorithm selected lexical items. This time, the next node to be visited at any iteration is the one which minimises the increase in total distance between *perspectives represented in a description*. A *perspective* $\mathcal{P}$ is represented

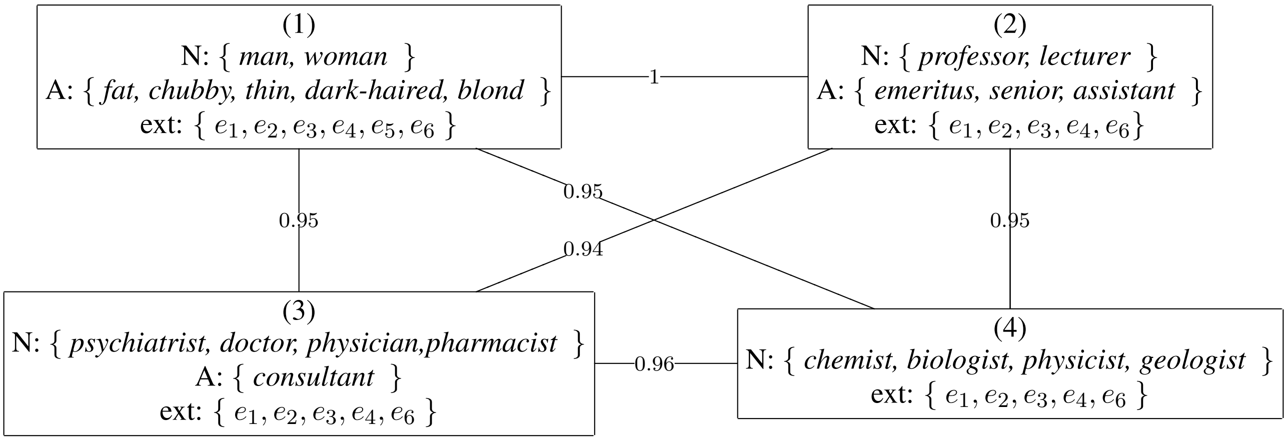

(1)
N: { man, woman }
A: { fat, chubby, thin, dark-haired, blond }
ext: { $e_1, e_2, e_3, e_4, e_5, e_6$ }
(2)
N: { professor, lecturer }
A: { emeritus, senior, assistant }
ext: { e_1, e_2, e_3, e_4, e_6 }
1
0.95
0.95
0.94
0.95
(3)
N: { psychiatrist, doctor, physician,pharmacist }
A: { consultant }
ext: { e_1, e_2, e_3, e_4, e_6 }
0.96
(4)
N: { chemist, biologist, physicist, geologist }
ext: { e_1, e_2, e_3, e_4, e_6 }

Figure 7.4: Perspective graph generated from the KB in Table 7.2

in the description if there is at least one lexical item in the description which is also in $\mathcal{P}$.

If perspectives contain items which are similar enough to 'go together', then it is safe to select any combination of items from the same perspective. Thus, rather than select nouns and adjectives on the basis of the overall distance in a description, the algorithm now proceeds incrementally, first through the set of nouns, then through the adjectives in the cluster, adding lexical items to the description whenever they have contrastive value. This results in a relativised interpretation of the weak local conceptual coherence constraint of Definition 12. Rather than on the basis of distance between nouns in a description, the new procedure views conceptual coherence in terms of the distance between perspectives represented in that description. The greedy procedure selects the cluster to be visited next. Within a particular cluster, the choice of a lexical item is carried out by a version of the Incremental Algorithm (Dale and Reiter, 1995) which simply orders nouns before adjectives.

Each cluster in Figure 7.4 also indicates the *extension* of the perspective; this is the set of domain entities of which the lexical items are true. Thus, a perspective is formally equivalent to a disjunction of all its elements; its extension is the union of the extension of all the lexical items in it:

$$ext(\mathcal{P}) = \bigcup_{l \in \mathcal{P}} [\![\text{ SEM}(l)]\!] \tag{7.20}$$

This new algorithm will still need to determine the root node for search. Because the information has been reduced, and related items are clustered together, it is relatively safe to calculate the discriminatory power of a cluster, rather than a single noun, assuming that because clusters contain several items, starting from the most discriminatory increases the likelihood that all referents be described from the same perspective. The discriminatory power of a cluster, $disc(\mathcal{P})$, is estimated by taking into account the referents in R that are also in its extension, and the distractors excluded:

$$disc(\mathcal{P}) = \frac{|ext(\mathcal{P}) \cap R| \; + |ext(\mathcal{P}) - C|}{|ext(\mathcal{P})|} \tag{7.21}$$

7.6.2 An example

To give an example of how this algorithm works, I return to the earlier example of generating a reference to $R = \{e_1, e_3\}$ in Table 7.2. The first task performed by the algorithm is to identify the perspective with the highest discriminatory power. In Figure 7.4, this is cluster 4, which contains *biologist* and *physicist*. The algorithm will not be able to distinguish the referents from this cluster alone. The cluster contains no adjectives, because all adjectives were clustered elsewhere. Hence, after traversing the set of nouns, the description consists of two DFs, each with a value of T_{DF} (*physicist* and *biologist* respectively), but an empty set M_{DF}. After this, the update procedure will remove the DF corresponding to e_3, because *physicist* is distinguishing for this referent.

The algorithm has to move to the next cluster, the one that has the shortest distance from the cluster represented in the description. In this case, there is a tie between clusters 1 and 2, which are at the same distance from cluster 4. Suppose cluster 1 is selected. This results in a dead end because none of the lexical items are true of the referents. Since this perspective is not represented in the description (and is not in $Nodes$), the next cluster visited will be the one containing *professor* and *lecturer* . The algorithm again begins by considering nouns. The only remaining referent to distinguish is e_1, and, while *professor* is true of it, it does not remove any distractors; hence $contrastive()$ returns false for this item. (The only other entity which is a biologist in the domain, namely e_7, is also a professor.) Therefore, adjectives are considered next. This time, *senior* does the trick, and the description returned is *the senior biologist and the physicist*. This description is identical to the one returned by the first version of the greedy algorithm without clustering (§7.5.2).

This procedure therefore returns an identical description to the one without information reduction. Nevertheless, there is the possibility of the description generated being more overspecified than the one returned by the previous algorithm. This is especially true because the procedure considers all nouns in a perspective first, then all adjectives. Thus, were another entity present in this domain, who was a biologist but not a professor, the latter lexical item would have been included, still requiring *senior* later to exclude e_7, who is an assistant. The result would be *the senior professor who is a biologist*. The results from previous chapters actually suggested that overspecification, even with plurals, is not viewed as problematic by human authors, if properties have sufficiently high codability to be included. Perhaps coherence is also a reason to include properties which are not strictly contrastive; indeed, this is what the experiment in §6.8 (p. 200) would suggest.

7.7 Activating the lexicon

The final computational model I consider to solve the problem of conceptual coherence is a priming-based model. Rather than explicitly take perspectives into account, this model is closer to the original greedy procedure introduced in §7.5, in that it considers similarity locally, between nodes of the Noun Graph, and conceives of the local conceptual coherence constraint as a property whose underlying cause is priming.

This model bears some relationship to previous work in cognitive modelling in which the notion of spreading activation is used to explain the ease of retrieval of information from long-term memory (e.g. Anderson et al., 1995; Anderson, 1996). Some computational psycholinguistic models of lexical retrieval fall broadly within this paradigm. Such models have converged on a view of the mental lexicon as a multi-tier associative memory structure (Dell, 1986; Roelofs, 1992; Levelt et al., 1999; Roelofs, 2000), in which lexical items are 'activated' and activation spreads to neighbouring items. The models proposed are multi-tier in the sense that links between lexical items, whereby activation spreads, exist at various levels, from the conceptual 'meaning' level to the phonological. Retrieval of an item during language production activates other lexical items, and the strength of the activation is a function of their distance from the selected item. Activation in these models can occur because a selected item shares syllabic structure with some of its neighbours (phonological level), but can also occur when items are related by virtue of their meaning. The models in the psycholinguistic literature differ on the details of the directionality of links between lexical items, and the precise locus of various priming effects (see Levelt, 1989, 1999, for a review). Nevertheless, these models are interesting because they make the prediction that the use of an item will prime the use of semantically related items, making them easier to put in short-term memory. Semantic priming has been observed in many laboratory settings (Meyer et al., 1998; Damian, 2000; Vigliocco et al., 2002; Damian et al., 2001).

Unlike the models of Dell (1986) and Roelofs (2000), the computational model proposed in this section is not intended to be a cognitive model; however, it has some kinship with these psycholinguistic models because it uses the idea of spreading activation at the semantic level. It should be emphasised, however, that the statistical definition of semantic relatedness used here is not shared with the views incorporated in these models, which in any case have tended not to give a detailed account of how association at the conceptual or meaning level is determined.

In the new priming-based algorithm, lexical items are active objects, and the Lexicon graph functions like an associative memory in that activation spreads from one item to another as a function of the similarity between items. If an item is selected, then it is strengthened (because it is in working memory, as it were), and this additional strength means that it boosts its neighbours more. To take an example from the Lexicon graph in Figure 7.2, the use of the word *professor* primes its neighbours in semantic space, by spreading activation. The nearer, semantically, a lexical item is, the higher the 'boost' it receives. Thus, *lecturer*, which is the closest item to *professor* in the graph, is the lexical item that is most highly primed once *professor* is selected. On the other hand, *physician* also receives a boost, but this is weaker, since the similarity between *professor* and *physician* is 0.094 ($\delta \approx .92$), compared to 0.145 ($\delta \approx .87$) for $\langle$*professor, lecturer*$\rangle$. There is, however, another factor: if *professor* is selected, then it is highly active in short-term memory; therefore, its activation will result in a stronger boost to nearby items.

I begin by giving the equation according to which activation spreads between lexical items. This is based on two assumptions:

1. Every lexical item has a strength $s(l)$;
2. Activation from an item l_1 to an item l_2 spreads if, and only if, $\langle l_1, l_2\rangle$ is an undirected, weighted edge or $\langle l_1, l_2\rangle$ is a directed weighted edge emanating from l_1.

Therefore, nouns spread activation to other nouns, but also to adjectives as a function of their collocational salience. The activation of an adjective also spreads among other adjectives; however, it does not spread from adjectives to nouns. This reflects the fact that selection of an adjective is dependent on a noun having been selected first, to which the adjective has a high collocational salience. To define the activation of a lexical item l, let $l_{\mathcal{IN}}$ be the set of nodes which are connected to l in the graph by edges that respect the restriction in (2) above. In case l is a noun, $l_{\mathcal{IN}}$

consists of all the edges that connect it to the other nouns in the nominal part of the lexicon. In case l is an adjective, $l_{\mathcal{IN}}$ contains only those edges connecting nouns to l. The activation of l, $act(l)$ is defined as follows:

$$act(l) = s(l) + \frac{\sum_{l' \in l_{\mathcal{IN}}} act(l')^{-\delta(l',l)}}{|l_{\mathcal{IN}} - 1|} \tag{7.22}$$

By this definition, the activation of a lexical item l is a function of (a) its strength, and (b) the incoming activation from neighbouring items (all edges incident on l in the graph), exponentiated by the (negative) distance that those items have from l. The amount of activation spread between lexical items therefore decreases exponentially as a function of the distance between them. Note that $act(l)$ is scaled by the number of items in $l_{\mathcal{IN}}$.

To continue with the previous example, suppose *professor* is selected. Assume, furthermore, that at state 0, before anything has been selected, every lexical item has unit strength. The selection of *professor* increases its strength by a constant k. Suppose k is 1. Then the activation of *professor* has now gone up to 2 (ignoring incoming activation from other lexical items). The activation spread to the neighbouring item *lecturer* is $act(\textit{professor})^{-.87} = .54$. Assuming that *lecturer* had unit strength, its new activation level is 1.54. In contrast, the activation spread from *professor* to *physician* is $act(\textit{professor})^{-.87} = .52$, and its new activation level is 1.52. This example is a simplification. As the equation shows, the incoming activation from neighbouring lexical items is actually divided by the number of nodes from which a lexical item receives activation. This is a way of normalising the amount of activation a lexical item receives.

For the purposes of the work described here, I have assumed that the strength of all lexical items at the start of content determination is 1, and that the constant k by which strength increases is also 1. However, the parameters s and k make the model flexible enough to take other factors into account. For example, Experiment 2 in Chapter 6 (§6.5.4, p. 182) showed a significant effect of frequency, whereby people were more sensitive to the coherence of an NP if the nouns in it were highly frequent. To take such a factor into account, the strength of a lexical item could be defined as a function of the frequency of that item in a corpus.

7.7.1 How the model works

This model works by priming. A lexical item that is selected will make it more likely that highly similar items will also be selected. The way it has been implemented, activation spreads according to the assumptions made above, so that (a) selection of a noun will make similar nouns more active; (b) selection of a noun also activates adjectives to the extent that they are collocationally related to the noun; (c) activation also spreads among adjectives.

The machinery that was put in place in my earlier discussion of the greedy algorithm in §7.5 is all that is needed in addition to the modifications to how the lexicon is conceived. However, the next lexical item returned by $nextItem()$ in Algorithm 6 is always the one which is most highly activated. Otherwise, the category of the item is still decided on the basis of the priority of a fragment.

In the new model, it is no longer necessary to start from a root node in the Noun Graph that is the most discriminatory. Rather, the root node for search is already the most highly active node. Since all nodes start from unit strength, the most highly active node will be the one that has the strongest incoming activation; therefore, it is also the one whose links to other nodes are strongest on average. This makes the maximisation of coherence more likely, as the starting point for search will always be the point of strongest convergence of similarities between nouns.

In the graph in Figure 7.2, *physicist* will be the strongest candidate for selection at the first iteration of the algorithm, because it is strongly linked to *geologist*, *biologist* and *chemist*. Therefore, in the case of a reference to $R = \{e_1, e_3\}$, it is this item that is selected first. Activation spreads to all nouns and adjectives from this item. The strongest link from *physicist* in the partial graph in the Figure is to *biologist*. This is, coincidentally, also a noun that has some discriminatory value for one of the referents. The point of the example, however, is to illustrate that this model is closer to the initial motivation for the algorithms presented here, insofar as it is a purely coherence-driven model, which does not include consideration of discriminatory value until a lexical item has been retrieved from the lexicon.

7.8 Discussion

The discussion of algorithms in this chapter took off from a description of a greedy model based on the graph-theoretic definition of the Lexicon. The two models proposed later, though partially motivated by computational considerations, were introduced as two alternative interpretations of the same class of phenomena. It is therefore useful to discuss these in the light of the empirical and theoretical work of the previous chapter.

In this work, what I have called a 'conceptual perspective' or a 'coherent conceptual cover of a set' is entirely defined on the basis of lexical information. To be sure, this was in part motivated by the work of authors such as E. Clark, whose Principle of Contrast (Clark, 1997a), stipulating that no two words are semantically *or* pragmatically identical, is stated at the lexical level, under the assumption that the acquisition of lexical items by speakers involves more than a memorisation of form-meaning pairings. Clark contends that factors such as the context of use, the communicative task, and the interlocutor all affect the way lexical items are learned. The corpus-derived definition of similarity used here was argued to reflect some of these factors insofar as they are reflected in language use. Under this definition, a pair of words w_1 and w_2 are similar the more they tend to co-occur with the same words in the same grammatical environments. This rest of this section briefly discusses the differences between the two models described in §7.6 and §7.7, and also considers the role of 'global' constraints in perspective-taking in reference.

The conceptual clustering model incorporates an explicit model of a conceptual or lexical perspective based on lexical similarity, making it the starting point for Content Determination. The spreading activation model, by contrast, works at a different level, and explains conceptual coherence, at least in the 'local' version that has been the focus of this work, rather as psycholinguists have explained tip-of-the-tongue phenomena or semantic activation and lexical retrieval in laboratory situations (Meyer, 1996; Meyer et al., 1998). Under this model, a description is perceived as coherent because a listener is primed to expect certain words to follow others. The same explanation carries over to the speaker's role (which is what the algorithm explicitly models): the use of a word primes related words, so that the more strongly activated a lexical item is by virtue of the selection of one of its semantic neighbours, the easier it is to retrieve. Therefore, in this model, conceptual coherence and the availability of a unified conceptual cover for a set are viewed as an ‘emergent’ property of the output: the reader/listener can infer the perspective taken on a set because she possesses the same lexical knowledge as the speaker and this knowledge includes the kinds of similarity relationships between lexical items that arise from the contexts in which words are used. In spite of their differences, both models are arguably ‘bottom-up’, insofar as their starting point is similarity relationships between lexical items. A more top-down perspective is conceivable, one which would incorporate global constraints such as communicative intention and relevance, and which would have more in common with pragmatically-oriented explanations for coherence, such as those offered by Kronfeld (1989) and Aloni (2002).

Consider a version of the algorithms presented here which took global constraints into account. Such a model would find perspectives (or select words) based not only on lexical similarity relations, but also by considering what things pertain to which communicative situations, how a given communicative intention can be satisfied by selecting one way of categorising referents rather than others, and so on. It would also be better suited to make an informed decision about when and whether to shift the perspective taken on a set of referents, given a new communicative intention. Such perspective shifts were exemplified in §6.2 (p. 167): in a text which initially described two entities using proper names, and later referred to them as *the master and the pupil* (cf. example 6.3), the decision to shift the lexical perspective may be made in order to meet a specific communicative goal. Accounting for these phenomena arguably requires a more knowledge-rich approach. To what extent does a model based exclusively on lexical similarity incorporate such global constraints?

I believe the answer to this question depends on how lexical similarity is defined. The definition used here, which reflects language use, was adopted because simpler definitions based on taxonomic knowledge were shown to be poorer predictors of people’s preferences. Clearly, any such result is partially dependent on the taxonomy used. However, there is also another motivation for using distributional similarity, already hinted at in the beginning of this section, and discussed more thoroughly in (§6.4, p. 176). Semantic regularities found through the analysis of large samples of naturally-occurring discourse will to some degree reflect the common situational and communicative constraints under which such discourses were produced. This is not to argue that pragmatic constraints are completely inferrable from statistical analysis of data; rather, it is to affirm that because discourse is grounded in real-world contexts, statistical regularities offer a window onto those contexts. This view is not too far from the view of meaning espoused by H. Clark (1991) and E. Clark (1997a). It is also supported by studies showing evidence of lexical priming in discourse understanding (e.g. Traxler et al., 2000; Jescheniak et al., 2005) and other studies which show that lexical similarity or relatedness is a determinant of discourse coherence (e.g. Morris and Hirst, 1991; Foltz et al., 1998).

This body of work, together with the evidence gleaned in the previous chapter, suggests that a definition of conceptual coherence based on the lexicon tells an important part of the story. However, this story cannot be assumed to be complete. It is only by incorporating more ‘top-down’ intentional constraints with lexical knowledge that a full account of conceptual perspective in reference can be given.

7.9 Summary and conclusions

This chapter proposed a number of models for achieving local conceptual coherence, all of which started out from a greedy heuristic that was bound to the representation of lexical information. The algorithms were focused primarily on content determination. One of the main innovations was to introduce lexicalisation as a prior step, and to structure lexical information in such a way that retrieval of lexical items could make use of similarity relationships between them. The work presented in this chapter points towards at least two further directions for future research. The first, already raised in §7.8, has to do with global coherence, whereby pragmatic and intentional factors play a role in determining which perspective to take on a referent at the outset. The second has to do with realisation, and how this is interleaved with content determination. Already, the work presented here does away with one standard assumption in GRE, namely that properties are what populate the search space of a GRE algorithm. As with the algorithm of Chapter 5, some aggregation was also performed. Interleaving this with incremental realisation seems a natural way to proceed. As discussed in §2.7.7 (p. 63), this approach has begun to arouse the interest of researchers in the area (e.g. Stone and Webber, 1998; Krahmer and Theune, 2002; Horacek, 2004).

Lexicalisation however raises novel problems for content determination. These are issues that a purely semantically-driven approach may ignore, but which become more topical in a framework such as the current one. The first problem has to do with lexical ambiguity. A specific case arose with the word *chemist*, which has both a ‘research chemist’ and a ‘practitioner’ reading. I tentatively proposed a Firthian view on this matter, arguing that words are disambiguated in the context of similar lexical items. This, however, remains a topic that is open to further research.

Another issue is related to how purely lexically-driven a content determination algorithm can be. The framework used here took into account the semantics of lexical items, because these were property-word pairings from the start. Further complications arise with non-intersective modifiers and multi-word expressions. The treatment of modification in the preceding sections largely ignored the differences between words like *thin* and words like *senior*, when modifying a noun such as *lecturer*. Arguably, *senior lecturer* is a multi-word expression, and denotes a sub-sort of *lecturer*. On the other hand, taking collocational salience into account is one way of dealing with the fact that *senior lecturer* is a frequent combination. Purely word-based accounts will run into problems with non-intersectivity. One such case is *emeritus*, which is a highly salient modifier of *professor*, but is attributive rather than intersective.[6] My treatment of these has been largely in line with most GRE work, where extensionality determines the denotation of a combination of properties. A possible alternative is to extend the distributional account of similarity to multi-word expressions, although such expressions are notoriously difficult to determine with accuracy from raw data. Another possibility is to couple the distributional information at the word level, with an ontological or taxonomic support. This remains an under-explored area in GRE (but cf. Croitoru and van Deemter, 2007, for a recent proposal that distinguishes ontological knowledge from domain instantiation). However, it also holds promise for broadening the notion of conceptual perspective used here to encompass such questions as relevance to a topic.

[6] In the sense that *emeritus professor* does not denote the intersection of things which are emeritus and things which are professors.

Chapter 8

Conclusion

8.1 Main contributions of the thesis

This thesis began by posing the question of what it is that makes a reference to an object adequate. Its main concern was twofold. First, it sought to find a psycholinguistically-motivated definition of adequacy. Second, its focus was on a generalisation of the standard problem of generating references, to arbitrary sets of objects. Plurality has long been something of a *bête noir* in GRE. The progress made in the area with singular reference, on questions of computational efficiency and content selection, seems to have been lost when plurals were brought into the picture. This state of affairs was exacerbated by a lack of psycholinguistic research on plurals, and by a tendency in the field of GRE itself to stop short of evaluating its models, or even to find empirical evidence to motivate them, where such evidence is lacking.

It is from the latter point that the present work took off. The first part of the thesis presented an exhaustive empirical analysis of references in a corpus, which also included plurals. In this part, a focal point was an evaluation of algorithms that characterise the state of the art in GRE. The results shed new light on these algorithms. For example, the 'gold standard' content determination procedure, the Incremental Algorithm Dale and Reiter (1995), was found to be highly dependent on a preset parameter – the preference order – and can run into problems, both when compared to speakers who are not self-consistent in their descriptive strategies, and when deployed in situations in which no a priori preference order can be discerned. With respect to the latter case, some new results by van der Sluis et al. (2007) suggest that determining such a parameter becomes even more non-trivial when the domain is even slightly more complex (and less familiar from previous psycholinguistic work) than the ones considered here.

The empirical work presented in this part of the thesis was also intended to make a methodological contribution. While corpora are now standard tools in NLP research, the corpus used here emphasised issues of balance and transparency, which were achieved by the use of a psycholinguistic experimental methodology for data collection. This enabled precise hypotheses to be formulated in relation to what content people use in their descriptions (as distinct from how they realise that content), and to also predict, based on these results, how algorithms would perform when exposed to exactly the same domains as corpus authors had been.

With respect to the original question – that is, the generalisation of algorithms to deal with sets rather than individuals – current approaches were shown to be severely impoverished in their performance compared to people. Why is this?

It seems that extensions of the GRE problem definition to plurals did not always take the most scientifically parsimonious route in generalising the problem. One of the first systematic approaches to plurality (van Deemter, 2002) showed that an existing algorithm could be generalised to achieve logical completeness, and therefore guarantee a logically correct outcome whenever one existed. This outcome, however, was shown to sometimes run counter to intuitions of what makes an adequate description. Subsequent approaches sometimes returned to a principle of adequacy based on brevity, which had been falsified for the case of singular reference over approximately three decades of psycholinguistic research. That people are not 'strictly Gricean' in formulating references (that is, they do not generate the briefest possible description) is usually explained with reference to automatic processes involved in language production. Now, plurals could well represent a more difficult case, one in which such processes are curbed in order to satisfy other constraints, such as logical simplicity or avoidance of excessive redundancy. Yet the more parsimonious approach to the problem would be to begin from the assumption that the same automatic, incremental production processes will determine the outcome of people's descriptions.

The empirical analysis of plurals in this thesis sought to substantiate this hypothesis, and showed that overspecification is in fact the norm, even with sets of objects. Crucially, however, independently motivated principles related to perception and conceptualisation also play an important role. One of the core hypotheses of this work has been that if, rather than a single entity, the focus of the referential intention is a set, then its conceptualisation will be facilitated if its elements are similar, whether this similarity is perceptual (for instance, all elements are of the same colour) or conceptual (for instance, all elements can be described from a related perspective). Besides the corpus data, a number of psycholinguistic experiments were reported to substantiate this claim, which was formulated under a Principle of Similarity.

From an algorithmic point of view, the issues investigated here posed a number of challenges. One was to maintain tractability, something which has been shown to be lost when current algorithms are generalised. Another challenge lay in an extension of the content determination problem in GRE, in two directions. The first concerned the interaction of *form* and *content*. It turned out that, in addition to similarity, people's descriptions of sets tend to be strongly influenced by how the objects they are referring to are categorised. This actually ties in well with the notion that a plurality which is the object of a referential intention is a conceptual gestalt, whose core is the class or category to which its entities belong. Where entities belong to different categories, people's descriptions evince a partitioning strategy, which interacts with the Principle of Similarity, sometimes increasing the redundancy of content in the descriptions thereby produced. This evidence led to the formulation of a Principle of Category-driven Reference. As a result, an algorithm was proposed which took a partitioning strategy, and also attempted to maximise similarity. This was achieved in part by including a statistical model of what attributes people are likely to include to maximise the coherence of their description, even when these attributes do not contribute to distinguishing a set from its distractors. This algorithm is also the first computational procedure for plural reference to have been evaluated empirically, and was shown to be computationally tractable.

Another extension of the remit of GRE was to encompass some aspects of other microplanning tasks, particularly aggregation and lexicalisation. The former was motivated by a need to balance the syntactic complexity of plural descriptions with the other principles that motivated people's content determination decisions. The latter was introduced as a way of addressing another core finding, namely, that people conceptualise sets under the same perspective as far as possible, and their tendency to do so can be partially predicted from the distributional similarity of the words they use. The algorithms discussed later in the thesis were based on an experimentally-motivated generalisation of the Principle of Similarity to one of Conceptual Coherence. This principle says that in referring to multiple entities, the same perspective should be taken on the set as far as possible. Perspective was viewed primarily from the lexical angle, based on prior work in pragmatics and psycholinguistics. For this reason, the algorithms subsequently proposed, while maintaining the core partitioning strategy developed earlier, made content determination lexically-driven.

A lexically-driven notion of perspective is of course only part of the story. The determinants of a coherent perspective on a set (or indeed on a single entity) include contextual factors and world knowledge. However, lexical choice plays a very important role, and a comparison of the family of algorithms proposed to deal with it against a Brevity-oriented model, showed that lexical coherence as interpreted by the algorithms yields descriptions that people tend to find better.

8.2 Remarks on methodology

The methodology used here can be summarised as follows. Hypothesis-formation was generally followed by experimental testing, and corpus analysis was used to find positive evidence for various phenomena. If such evidence converged, then it was applied in the task of algorithmic modelling. In the case of Chapters 3, 4 and 5, the experimental and corpus-based methodology came together to some extent, because the semantically transparent corpus used in those chapters was constructed using a controlled experiment. Moreover, that corpus was specifically designed to address the semantically intensive task of content determination.

I believe that this method has a lot to be said in its favour. However, when the focus is Content Determination, experiments are often tricky. To take an example, the Content Determination experiment of §6.7 (p. 196), as well as the experimental comparison of the Brevity and Coherence models in §6.8 (p. 200) gave people choices of content using a linguistic modality. This seems to be the only alternative in such experiments. Clearly, a potential criticism of an approach that uses these experiments to make inferences for a *semantic* task is that the results were already realised, so that the inference of constraints on content determination takes place at one remove. On the other hand, an experimental methodology allows falsification, which can sometimes lead to the positing of further questions. This was the case with the results of the Content Determination experiment in relation to modifiers, for example.

Perhaps the best way to proceed in such tasks is to attempt to strike a balance between experimental and corpus-based work; while the former is an invaluable source of evidence for specific hypotheses, the latter often yields statistically interesting insights into the workings of language. Some of the work in this thesis tried to bridge the two, by using corpus-derived estimates of statistical similarity in psycholinguistic experiments, before incorporating them into GRE algorithms.

8.3 Directions for future research

The empirical investigations and the algorithms presented in the preceding chapters open several avenues for future work. I have occasionally discussed these in other sections; here, I will point out some possible extensions of the model of reference based on similarity and coherence. Three stand out in particular.

The first promising research direction is plural anaphora. Although anaphora has been the focus of some work in GRE in the past, plural anaphors have received far less attention. The Conceptual Coherence Hypothesis posited in Chapter 6 was in part motivated by work in psycholinguistics that has shown that pluralities in discourse are gestalts ('more than the sum of their parts') and that similarity or conceptual relatedness plays a role in how easy they are to represent. The prediction that pluralities have separate status from their elements, and that this is partially contingent on similarity, raises interesting questions. For example, how should a a context-sensitive GRE algorithm (e.g. Krahmer and Theune, 2002) deal with salient plural referents? Such referents contain more than one entity; are all elements of a plurality equally salient or does the plurality as a whole have a separate status from those elements? The work presented here would most naturally favour the second alternative, since part of it was motivated by the hypothesis that pluralities are gestalts (and hence more than the sum of their parts). The results on Local Conceptual Coherence may also be relevant to anaphora resolution, where the similarity between potential antecedents may help to resolve a plural anaphor.

A second area of research is similarity and coherence of conceptualisation *across noun phrases*. Adopting a conceptual perspective on a set (including a singleton) may bias subsequent references, not only to the same set, but also to entities in the same discourse context. On the other hand, a change in perspective is also an important pragmatic cue. Some work, for example by Traxler et al. (2000); Jescheniak et al. (2005), has shown evidence for lexical priming within a discourse. A study by Foltz et al. (1998) gave strong evidence that a statistical definition of similarity (using Latent Semantic Analysis in this case) was a good predictor of the coherence of segments of extended discourses.

In this connection, the third issue arises, which has to do with the role of knowledge, both ontological and situational, in taking perspective on a set or individual. Several questions have not been addressed here, including the question of relevance, which has been flagged by authors such as Kronfeld (1989) and Aloni (2002) as an important factor in determining what perspective is taken in reference or question answering. The challenge for future work is to combine a top-down perspective-taking mechanism, based perhaps on structures representing world knowledge, with the bottom-up lexical forces employed in the present study.

8.4 Issues for Natural Language Generation

In recent years, NLG has become increasingly oriented towards empirical work, especially where evaluation of systems is concerned. This thesis bears much in common with this trend. However, the use of an experimental methodology distinguishes this work from much of the work in the field, which is dominated by a corpus-based approach. Clearly, large balanced corpora, or small domain-specific collections, are invaluable research tools for NLG; however, they are a source of exclusively positive evidence. Moreover, not all NLG tasks benefit equally from linguistic corpora, unless these are semantically transparent. I believe that work in NLG would also benefit from a methodology which borrowed techniques from kindred fields, such as psycholinguistics, which permit hypotheses to be addressed (and falsified) in a more focused way. However, such work tends to be costly and time-consuming, and the costs may well outweigh the engineering benefits in the long run. Nevertheless, if NLG is, apart from an engineering challenge, also a contributor to cognitive science, such work may be valuable on other grounds.

Another facet of current work in NLG that has been alluded to at various points in this work is the trend towards more global approaches to the generation task, interleaving semantic tasks such as content determination, with realisation and aggregation. This seems to be a promising way forward, even for a semantically-intensive area like GRE, because constraints on the generation of language are defined at multiple levels, and frequently interact.

If current trends are anything to go by, future NLG systems will be based on more thorough empirical grounding, and will also take a more holistic approach to the language generation task. Perhaps some of the work presented here may be viewed as falling within the scope of these current trends.

Appendix A

Instructions given to authors in the corpus

Below is the full text of instructions given to participants in the experiment for the construction of the corpus described in Chapter 3.

A.1 Instructions common to all versions

In this experiment, we are trying to evaluate the performance of a computer program that understands English. You will be given a task in which you describe and identify objects for the computer. The computer will then try to interpret your description. Here is what the task entails:

You will be shown a number of scenarios. In each one you'll see pictures of **furniture** or of **people**. Some of them will be surrounded by a red border. Your task is to answer the question **Which objects are surrounded by a red border?** Write the answer in the box provided, as though you were speaking to a normal person.

A.2 Condition 1: +FC+LOC

Each time you do this, click the *submit* button. The program will then try to figure out which objects you mean, and remove them from the screen. It can "see" exactly the same pictures as you **in exactly the same position**.

Our program will eventually be used in situations where it is crucial that it understands descriptions accurately with no option to correct mistakes. Therefore, in this experiment, if it misunderstands your description, you will not get the chance to revise it. Moreover, you will not be able to use the *Back* or *Refresh* buttons on your broswer to describe the same objects again.

A.3 Conditoin 2: +FC−LOC

Each time you do this, click the *submit* button. The program will then try to figure out which objects you mean, and remove them from the screen. It can "see" exactly the same pictures as you. However, **the position of the pictures has been jumbled up in its version, so they don't appear in the same position as in your version.**

Our program will eventually be used in situations where it is crucial that it understands descriptions accurately, with no option to correct mistakes. Therefore, in this experiment, if it misunderstands your description, you will not get the chance to revise it. Moreover, you will not be able to use the *Back* or *Refresh* buttons on your broswer to describe the same objects again.

A.4 Condition 3: −FC−LOC

Each time you do this, click the *submit* button. The program will then try to figure out which objects you mean, and remove them from the screen. It can ”see” exactly the same pictures as you. However, **the position of the pictures has been jumbled up in its version, so they don’t appear in the same position as in your version.**

If the computer misunderstands your description and removes the wrong objects, you can point out the right objects for it, by clicking on the pictures with the red borders.

Appendix B

Materials for the Magnitude Estimation experiments reported in Chapter 6

B.1 Instructions

The purpose of this exercise is to get you to judge the acceptability of some English sentences or phrases. These judgments should be based entirely on your intuitions as a speaker of English. There are no right or wrong answers.

You will see a series of sentences or phrases on the screen. Some will seem perfectly okay to you, but others will not. What we're after is whether the sentence sounds natural to you or not. This means that you should judge **whether you are likely to hear or use the phrase in some situation**.

You will be asked to make your judgments by comparing each phrase with an **anchor phrase**[1]. This will be the first phrase that you'll see. You will first judge the anchor and then compare each sentence to it. You will be asked to make your judgments in either of two ways:

- **Numerically**: Sometimes you will be asked to give a number to the phrase on a scale of your choice. The greater the number, the more acceptable the phrase is to you. When judging a phrase numerically:

    1. you can use any range of positive numbers you like, including fractions or decimals if you wish
    2. you should not restrict your scale to an academic marking scale (e.g. from 1 to 10)
    3. you may not use negative (minus) numbers or zero

- **Visually**: Sometimes you will be asked to judge a phrase by moving a slider on a line. The further to the right you move the slider, the more acceptable the phrase is to you.

In the rest of this experiment, you will first be given some practise with numeric scales and sliders. Then, you will be asked to rate your anchor phrase, both numerically and visually. Remember, every other phrase you see will be compared to this one.

After that, you're ready to go. With each sentence or phrase, you'll be shown your original anchor phrase. Just judge the new item in proportion to the anchor. For example, if you like the new phrase twice as much as the anchor phrase, give it a number twice the size, or move the slider

[1] *Anchor phrase* was the non-technical term adopted to refer to the *modulus* item in the experiment

twice the distance. Do not spend too much time thinking about sentences; what's important to us is your intuition.

B.2 Materials used for Experiment 2 (§6.5.4)

These phrases represent combinations of the three factors:

1. Frequency (FR): High/Medium/Low;
2. Distributional Similarity (DS): High/Low
3. Ontological Relatedness (OR): High/Low

FR	DS	OR	**Phrase**
High	High	High	the leader and the chairman
High	High	Low	the manager and the council
High	Low	High	the department and the resource
High	Low	Low	the garden and the police
Medium	High	High	the guitar and the piano
Medium	High	Low	the essay and the publication
Medium	Low	High	the lorry and the satellite
Medium	Low	Low	the printer and the coin
Low	High	High	the tumor and the ulcer
Low	High	Low	the rug and the poster
Low	Low	High	the tutor and the suspect
Low	Low	Low	the staircase and the truck

B.3 Materials used for Experiment 3 (§6.5.5)

These phrases represent combinations of the two factors:

1. Distributional Similarity (DS): High/ Low
2. Ontological Relatedness (OR): Homogeneous Animate/ Homogeneous Inanimate/ Heterogeneous

Each combination is in both Sentence and Phrase form, corrsponding to the Stimulus Type (ST) between-groups factor. Further, all conditions are represented by two phrases/sentences a in two versions A and B, which correspond to different judgment modalities.

DS	OR	Phrase	Sentential predicate
High	Animate	the secretary and the manager	were full-time
High	Inanimate	the table and the desk	were polished
High	Heterogeneous	the journalist and the newspaper	were British
High	Animate	the teacher and the student	were foreign
High	Inanimate	the bottle and the glass	were empty
High	Heterogeneous	the author and the novel	were popular
Low	Animate	the plumber and the waitress	were tall
Low	Inanimate	the carpet and the violin	were new
Low	Heterogeneous	the boy and the chair	were small
Low	Animate	the technician and the nun	were good
Low	Inanimate	the computer and the door	were ordinary
Low	Heterogeneous	the politician and the shoes	were Italian

Appendix C

Materials for the sentence continuation experiment reported in §6.7

The 16 discourses used in the Nominal and Modifier conditions are reproduced below. Similar pairs of nouns and/or adjectives are coindexed. Each discourse is followed by two continuations, used in the two versions of the experiment.

C.1 Nominal condition

Discourse 1

Three of the richest men in Europe were spotted by this newspaper last night, having dinner in a private suite at a London restaurant. All three men are millionaires, with a passion for fine arts and antiques.

(e_1) One of the men, a Rumanian, is a dealer$_i$.

(e_2) The second, a prince$_j$, is a collector$_i$.

(e_3) The third, a duke$_j$, is a bachelor.

C1 The XXXXXXXXXXXXXXXXX have both been in England for some time now, but the Rumanian was seen for the first time yesterday.

C2 The XXXXXXXXXXXXXXXXX have both been in England for some time now, but the bachelor was seen for the first time yesterday.

Discourse 2

The Gallery of the Artists' Cooperative reopened yesterday with a joint exhibition by an international group of artists. The theme of the exhibition is religious and cultural diversity. This newspaper interviewed three people at the opening.

(e_1) One of the people exhibiting is a Hindu$_i$. She is the curator.

(e_2) We also interviewed a Russian, who is a painter$_j$.

(e_3) Another interviewee is a Muslim$_i$. He is a sculptor$_j$.

C1 The XXXXXXXXXXXXXXXXX have both been involved in several awareness campaigns in the past, but the curator said this was the first time she was participating.

C2 The XXXXXXXXXXXXXXXXX have both been involved in several awareness campaigns in the past, but the Russian said this was the first time she was participating.

Discourse 3

Three men launched a community outreach project in a suburb of Glasgow, which aims to target minority ethnic groups, to help integrate them in the community.

(e_1) One of the men, a shopkeeper, is a businessman$_i$.

(e_2) Another member, a theologian$_j$, is a lecturer$_i$.

(e_3) Another, a clergyman$_j$, is an activist.

C1 The XXXXXXXXXXXXXXXXX had been planning this for many years before teaming up with the businessman, who will provide funding for the project.

C2 The XXXXXXXXXXXXXXXXX had been planning this for many years before teaming up with the lecturer, who is acting as an advisor.

Discourse 4

A prominent pharmaceutical company announced yesterday that three of its employees had been fired due to professional misconduct, following a scandal in which illegal substances were found in one of the company's products.

(e_1) One of the former employees is a technician, who was an adminstrator$_i$.

(e_2) The second employee to be fired is a chemist$_j$. He was a researcher.

(e_3) Finally, the company also fired a pharmacist$_j$. He was an assistant$_i$.

C1 The XXXXXXXXXXXXXXXXX had both been employees for several years, while the technician had only started a month ago.

C2 The XXXXXXXXXXXXXXXXX had both been employees for several years, while the researcher had only started a month ago.');

Discourse 5

The London Mayor yesterday announced the winners in the different categories of the Senior Citizen of the Year awards.

(e_1) One of the people who recieved a prize is a widow$_i$, who was a philanthropist.

(e_2) Special mention was given to a footballer$_j$. He used to be a striker.

(e_3) Another prestigious prize went to a pensioner$_i$, who used to be a referee$_j$.

C1 Both the XXXXXXXXXXXXXXXXX had won awards before, but this was the first time the philanthropist had won anything.

C2 Both the XXXXXXXXXXXXXXXXX had won awards before, but this was the first time the striker had won anything.

Discourse 6

Animal rights campaigners targeted the laboratory of a pharmaceutical company yesterday. After the attack, police questioned three people.

(e_1) The first person, a woman, is the president$_i$.

(e_2) The second person, an Irishman$_j$, is the secretary$_i$.

(e_3) The third, a Londoner$_j$, is a campaigner.

C1 Both the XXXXXXXXXXXXXXXXX have been involved in similar cases in the past, but this is the first time that the campaigner was suspected.

C2 Both the XXXXXXXXXXXXXXXXX have been involved in similar cases in the past, but this is the first time that the woman was involved.

Discourse 7

A fire broke out at a private clinic in Glasgow on Tuesday afternoon. The fire brigade arrived shortly after. Later, a fireman reported that only three people had been present at the time of the accident.

(e_1) One of them, a doctor$_i$, is an Englishman.

(e_2) Another, a trainee$_j$, is a nurse$_i$.

(e_3) Also involved is a Frenchman, who is a supervisor$_j$.

C1 Both the XXXXXXXXXXXXXXXXX escaped unhurt, but the Frenchman suffered minor injuries.

C2 Both the XXXXXXXXXXXXXXXXX escaped unhurt, but the Englishman suffered minor injuries.

Discourse 8

A university building was robbed last night. The police have detained three suspects for questioning, all of whom work or study at the university.

(e_1) One of them is a postgraduate$_i$. He is a physicist.

(e_2) Another is, a Greek$_j$, an undergraduate$_i$.

(e_3) Also among the suspects is a cleaner. He is an Italian$_j$.

C1 Both the XXXXXXXXXXXXXXXXX were held in custody, but the cleaner was released last night.

C2 Both the XXXXXXXXXXXXXXXXX were held in custody, but the physicist was released last night.

C.2 Modifier condition

Discourse 1

Hal wanted to buy a new computer, so he went to a computer superstore. There, the salesman showed him several models.

(e_1) One of them was a second-hand$_i$ computer. It was slow$_j$.

(e_2) Another was a brand-new$_i$, lightweight computer.

(e_3) He also saw a fast$_j$, American one.

C1 Since both the XXXXXXXXXXXXXXXXX were too big for his desk, he opted for the American one.

C2 Since both the XXXXXXXXXXXXXXXXX were too big for his desk, he opted for the lightweight model.

Discourse 2

Before selling his house, Dave decided to auction off some of the furniture. However, there were three vases he thought might be valuable, so he took them to an antique dealer for advice.

(e_1) One of them was an Oriental$_i$ marble$_j$ vase.

(e_2) Another one was a black vase, which was Persian$_i$.

(e_3) There was also a bronze$_j$ vase. It was valuable.

C1 Dave decided not to sell the XXXXXXXXXXXXXXXXX because he liked them both. He sold off the valuable vase for a lot of money.

C2 Dave decided not to sell the XXXXXXXXXXXXXXXXX because he liked them both. He sold off the black vase for a lot of money.

Discourse 3

Chris and Rachel wanted to buy a new coffee table, but there were so many in the furniture shop, they didn't know which to choose.

(e_1) One table they liked was narrow$_i$. It was antique$_j$.

(e_2) However, they also liked one which was made of mahogany. It was wide$_i$.

(e_3) Another option was a modern$_j$ table, which was low.

C1 The XXXXXXXXXXXXXXXXX were both too shiny, so they bought the low table.

C2 The XXXXXXXXXXXXXXXXX were both too shiny, so they bought the mahogany table.

Discourse 4

Joe was trying to clear up the mess in his young daughter's room. He'd stowed most of the toys away when he realised there were still a few dolls left lying about.

(e_1) There was a small rubber$_j$ doll.

- (e_2)] There was also a large$_i$ Asian doll.

(e_3) He also noticed a lifesize$_i$ wooden$_j$ doll.

C1 The XXXXXXXXXXXXXXXXX both went into the cupboard, while he put the Asian doll on the shelf.

C2 The XXXXXXXXXXXXXXXXX both went into the cupboard, while he put the lifesize doll on the shelf.

Discourse 5

The police raided the apartment of a suspected fraudster. They confiscated three folders which contained valuable evidence against him.

(e_1) One was thin$_i$ and made of paper$_j$.

(e_2) Another was made of plastic$_j$. It was stained.

(e_3) The third was a fat$_i$ brown folder.

C1 The XXXXXXXXXXXXXXXXX both contained information about the man's victims, but the stained folder contained correspondence.

C2 The XXXXXXXXXXXXXXXXX both contained information about the man's victims, but the brown folder contained correspondence.

Discourse 6

The company manager decided his office needed to be refurbished because the furniture was mismatched. He had three filing cabinets, none of which he really liked.

(e_1) One of them was a wooden$_i$ cabinet, which was old$_j$.

(e_2) He also had a metal$_i$ filing cabinet. It was full.

(e_3) Then there was a new$_j$ one, which was empty.

C1 He decided to change both the XXXXXXXXXXXXXXXXX, but kept the metal cabinet.

C2 He decided to change both the XXXXXXXXXXXXXXXXX, but kept the wooden one.

Discourse 7

The children had accompanied Pete on a shopping trip. To reward them for their good behaviour, he took them to a toy shop, where he told the salesman he wanted a ball for his children, but didn't want to spend too much.

(e_1) The salesman suggested a blue$_i$ leather ball.

(e_2) Pete also noticed a small$_j$ ball, made of plastic.

(e_3) Eventually, he settled on a large$_j$ green$_i$ ball.

C1 His children had hoped he'd buy both the XXXXXXXXXXXXXXXXX, because they didn't like the small ball.

C2 His children were hoping hed́ buy both the XXXXXXXXXXXXXXXXX because they didn't like the leather ball.

Discourse 8

John visited a bookshop on Saturday morning. He wanted to buy some books for his library. He came out of the bookshop carrying three books.

(e_1) He bought an old$_i$ thick$_j$ book.

(e_2) He also bought a recent$_i$ book, which was cheap.

(e_3) Having some money left, he also decided to purchase a thin$_j$, expensive book.

C1 He kept both the XXXXXXXXXXXXXXXXX but decided to give the recent book to his cousin.

C2 He kept both the XXXXXXXXXXXXXXXXX but decided to give the old book to his cousin.

Appendix D

Materials for the evaluation experiment reported in §6.8

Discourse 1

Three men who had been together at unviversity met in a pub for the first time in 5 years. One of them was Italian. He was a university lecturer, and had recently got married. The other, a university professor, was a famous writer. He was French. The third was also a university professor who had remained a bachelor. He was an Englishman. While they were talking, a waitress came to their table to take their orders.

$(+c, -m)$ The Italian, the Frenchman and the Englishman liked her immensely.

$(+c, +m)$ The lecturer and the professors liked her immensely.

$(-c, +m)$ The Italian and the bachelors liked her immensely.

$(-c, -m)$ The lecturer, the writer and the Englishman liked her immensely.

Discourse 2

A new exhibition opened at the Aberdeen Art Gallery, featuring works by Scottish artists. Among the works is a bronze sculpture of a female. There is also a painting, a portrait by a Glaswegian painter. He is also exhibiting a drawing of a landscape. All three works were especially commissioned for this occasion.

$(+c, -m)$ The sculpture, the painting and the drawing cost a lot of money.

$(+c, +m)$ The sculpture and the pictures cost a lot of money.

$(-c, +m)$ The female and the pictures cost a lot of money.

$(-c, -m)$ The female, the painting and the landscape cost a lot of money.

Disourse 3

A collection of old manuscripts was auctioned at Sotheby's. Three items attracted huge offers. One of them is a book, a biography of a composer. The second, a sailor's journal, was published in the form of a pamphlet. It is a record of a voyage. The third, another pamphlet, is an essay by Hume. All three were recent findings.

$(+c, -m)$ The biography, the journal and the essay were sold to a collector.

$(+c, +m)$ The book and the pamphlets were sold to a collector.

$(-c, +m)$ The biography and the pamphlets were sold to a collector.

$(-c, -m)$ The book, the record and the essay were sold to a collector.

Discourse 4

Police were investigating a murder. The body of a young woman had been found in a house. The police also found some drugs on the scene. Among these was a tablet. It was a powerful sedative. They also found a capsule containing antibiotic. Some syrup was spilled on the floor. The liquid turned out to be an antibiotic too.

$(+c, -m)$ The tablet, the capsule and the syrup were taken as evidence.

$(+c, +m)$ The sedative and the antibiotics were taken as evidence.

$(-c, +m)$ The pills and the liquid were taken as evidence.

$(-c, -m)$ The sedative, the capsule, and the liquid were taken as evidence.

Discourse 5

Three people have been nominated for the Nobel Peace Prize. One of them is a doctor who worked for the World Health Organisation. He is Greek. Another is Nigerian, a lawyer by profession. He is a politician. The third person to be nominated is a woman journalist, who acted as a war correspondent for many years. She is Nigerian too.

$(+c, -m)$ The doctor, the lawyer and the journalist were not available for an interview.

$(+c, +m)$ The Greek and the Nigerians were not available for an interview.

$(-c, +m)$ The doctor and the Nigerians were not available for an interview.

$(-c, -m)$ The politician, the Greek, and the woman were not available for an interview.

Discourse 6

A landmark industrial case is being heard in court. The defendant is a German manufacturer. He is being prosecuted by two individuals. One is an investor. He is a Spaniard. The other is a German retailer.

$(+c, -m)$ The manufacturer, the investor, and the retailer had all been collaborating on a project.

$(+c, +m)$ The prosecution and the defendant had all been collaborating on a project.

$(-c, +m)$ The investor and the Germans had all been collaborating on a project.

$(-c, -m)$ The Spaniard, the retailer, and the defendant had all been collaborating on a project.

Appendix E

Publications related to this thesis

The following is a list of the publications in which parts of this thesis, or material directly relevant to it, has appeared.

1. Gatt, A., and van Deemter K. (2005). Semantic similarity and the generation of referring expressions: A first report. *Proccedings of the 6th International Worskhop on Computational Semantics, IWCS-VI.*

2. Gatt, A. (2006). Structuring knowledge for reference generation: A clustering algorithm. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-06.*

3. Gatt, A. (2006). Generating collective spatial references. *Proceedings of the 28th Annual Conference of the Cognitive Science Society, CogSci-06.*

4. Gatt, A., and van Deemter, K. (2006a). Conceptual coherence in the Generation of Referring Expressions. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, COLING/ACL-06* [Main poster session].

5. Gatt, A., and van Deemter, K. (2006b). Conceptual coherence and the generation of plural references. *Proceedings of the Workshop on Coherence for Generation and Dialogue*, in conjunction with the *European Summer School in Logic Language and Information, ESSLLI-06*. [This is a revised version of Gatt and van Deemter 2006a].

6. van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. *Proceedings of the 4th International Conference on Natural Language Generation, INLG-06* [Special Session on Data Sharing and Evaluation].

7. Gatt, A., and van Deemter, K. (2007a). Lexical choice and conceptual perspective in the generation of plural referring expressions. To appear in: ¡i¿Journal of Logic, Language and Information (JoLLI)¡/i¿.

8. Gatt, A., and van Deemter, K. (2007b). Incremental generation of plural descriptions: Similarity and partitioning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-07*

9. Gatt, A., van der Sluis, I., and van Deemter, K. (2007a). Corpus-based evaluation of Referring Expressions Generation. Position paper at the *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, Arlington, Virigina.

10. Gatt, A., van der Sluis, I., and van Deemter, K. (2007b). Evaluating algorithms for the Generation of Referring Expressions using a balanced corpus. *Proceedings of the 11th European Workshop on Natural Language Generation, ENLG-07.*

11. van der Sluis, I., Gatt, A., and van Deemter, K. (2007). Evaluating algorithms for the Generation of Referring Expressions: Beyond toy domains. *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP-07.*

Bibliography

Albrecht, J. and Clifton, C. (1998). Accessing singular antecedents in conjoined phrases. *Memory and Cognition*, 26:599–610.

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, Stanford, Ca.

Amoia, M., Gardent, C., and Thater, S. (2002). Using set constraints to generate distinguishing descriptions. In *Proceedings of the 7th International Workshop on Natural Language Understanding and Logic Programming, NLULP'-02*.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34:351–366.

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, 51:355–365.

Anderson, J. R., John, B. E., Just, M. A., Carpented, P. A., Kieras, D. E., and Meyer, D. E. (1995). Production system models of complex cognition. In *Proceedings of the* 17^{th} *Annual Conference of the Cognitive Science Society*.

Appelt, D. (1985a). Planning english referring expressions. *Artificial Intelligence*, 26(1):1–33.

Appelt, D. (1987a). Bidirectional grammars and the design of natural language generation systems. In *Proceedings of the 3rd Conference on Theoretical Issues in Natural Language Processing, TINLAP-87*.

Appelt, D. and Kronfeld, A. (1987). A computational model of referring. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 640–647.

Appelt, D. E. (1985b). Some pragmatic issues in the planning of definite and indefinite noun phrases. In *Proceedings of the 3rd Annual Meeting of the Association for Computational Linguistics, ACL-85*.

Appelt, D. E. (1987b). Towards a plan-based theory of referring actions. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*. Nijhoff, Dordrecht.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Ariel, M. (2001). Accessibility theory: An overview. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text Representation: Linguistic and Psycholinguistic Aspects*. John Benjamins.

Arts, A. (2004). *Overspecification in Instructive Texts*. PhD thesis, Univiersity of Tilburg.

Bach, K. (1994). *Thought and reference.* Oxford University Press, Oxford.

Bach, K. (2005). The top 10 misconceptions about implicature. In Birner, E. and Ward, G., editors, *A Festschrift for Larry Horn*. John Benjamins, Amsterdam.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, CICLING-02.*

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Barry, C., Morrison, C. M., and Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures. *Quarterly Journal of Experimental Psychology*, 50A(3):560–585.

Barsalou, L. (1983). Ad hoc categories. *Memory and Cognition*, 11:211–227.

Barzilay, R. and Lapata, M. (2005). Collective content selection for concept-to-text generation. In *Proceedings of the joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, HLT/EMNLP-05.*

Barzilay, R. and Lapata, M. (2006). Aggregation via set partitioning for natural language generation. In *Proceedings of the joint Human Language Technology Conference and Meeting of the North American Chapter of the Association of Computational Linguistics, HLT/NAACL-06.*

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, A. and Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-06.*

Bierner, G. and Webber, B. L. (2000). Inference through alternative-set semantics. *Journal of Language and Computation.*, 1(2):259–274.

Bierwisch, M. (1989). The semantics of gradation. In Bierwisch, M. and Lang, E., editors, *Dimensional Adjectives.* Springer, Berlin.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *British Medical Journal*, 310:170.

Bock, K. and Levelt, W. (1994). Language production: Grammatical encoding. In Gernsbacher, M. A., editor, *Handbook of Psycholinguistics.* Academic Press, New York.

Bohnet, B. and Dale, R. (2005). Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05.*

Brennan, S. and Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

Brown-Schmidt, S., Campana, E., and Tanenhaus, M. K. (2002). Reference resolution in the wild: Online circumscription of referential domains in a natural interactive problem-solving task. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society.*

Bryant, D., Tversky, B., and Franklin., N. (1992). Internal and external spatial frameworks representing described scenes. *Journal of Memory and Language*, 31:74–98.

Cahill, L., Doran, C., Evans, R., Mellish, C., Paiva, D., Reape, M., Scott, D., and Tipper, N. (1999). In search of a reference architecture for NLG systems. In *Proceedings of the 8th European Workshop on Natural Language Generation, ENLG-99.*

Campana, E., Brown-Schmidt, S., and Tanenhaus, M. (2002). Reference resolution by human partners in a natural interactive problem-solving task. In *Proceedings of the 7th International*

Conference on Spoken Language Processing, ICSLP-02.

Carreiras, M. (1997). Plural pronouns and the representation of their antecedents. *European Journal of Cognitive Psychology*, 9(1):53–87.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., and Calrson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, 47:30–49.

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-05*.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, E. (1997a). Conceptual perspective and lexical choice in acquisition. *Cognition*, 64:1–37.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B., editor, *Mechanisms of language acquisition.* Lawrence Erlbaum, Hillsdale, N.J.

Clark, H. (1996). *Using language*. Cambridge University Press, Cambridge, UK.

Clark, H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. In Clark (1992). Reprinted in Clark (1992).

Clark, H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. In Clark (1992), pages 1–39. Reprinted in Clark (1992).

Clark, H. H. (1991). Words, the world and their possibilities. In Clark (1992). Reprinted in Clark (1992).

Clark, H. H. (1992). *Arenas of Language Use.* CSLI, Stanford, Ca.

Clark, H. H. (1997b). Dogmas of understanding. *Discourse Processes*, 23:567–598.

Cohen, W. (1996). Learning trees and rules with set-valued features. In *Proceedings of the 14th Conference of the American Association for Artificial Intelligence, AAAI-96.*

Corley, S., Corley, M., Keller, F., Crocker, M., and Trewin, S. (2001). Finding syntactic structure in unparsed corpora: The GSearch corpus query system. *Computers and the Humanities*, 35:81–94.

Cormen, T. H., Lesierson, C. E., Rivest, R. L., and Stein, C. (2003). *Introduction to Algorithms*. MIT Press, Cambridge, Ma.

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-07*.

Cruse, D. A. (1977). The pragmatics of lexical specificity. *Journal o f Linguistics*, 13:152–164.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge University Press., Cambridge.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

Dale, R. (2003). *One*-anaphora and the case for discourse-driven referring expression generation. In *Proceedings of the Australasian Language Technology Workshop.*

Dale, R. and Haddock, N. (1991). Generating referring expressions containing relations. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics.*

Dale, R. and Reiter, E. (1992). A fast algorithm for the generation of referring expressions. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-92*.

Dale, R. and Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

Dale, R. and Reiter, E. (1996). The role of the Gricean maxims in the generation of referring expressions. In *Proceedings of the AAAI-96 Spring Symposium on Computational Models of Conversational Implicature.*

Damian, M. F. (2000). Semantic negative priming in picture categorisation and naming. *Cognition*, 76:B45–B55.

Damian, M. F., Vigliocco, G., and Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, 81:B77–B86.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, 93:283–321.

Deutsch, W. and Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, 11:159–184.

Donnellan, K. (1966). Reference and definite descriptions. *Philosophical Review*, 75:281–304.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., and Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24:409–436.

Edmonds, P. G. (1994). Collaborating on reference to objects that are not mutually known. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94*.

Eikmeyer, H. J. and Ahlsèn, E. (1996). The cognitive process of referring to an object: A comparative study of german and swedish. In *Proceedings of the 16th Scandinavian Conference on Linguistics*.

Engelhardt, P. E., Bailey, K., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54:554–573.

Eschenbach, C., Habel, C., Herweg, M., and Rehkamper, K. (1989). Remarks on plural anaphora. In *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, EACL-89.*

Fodor, J. (1983). *The Modularity of Mind.* MIT Press, Cambridge, Ma.

Foltz, P., Kintsch, W., and Landauer, T. K. (1998). Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2 & 3):285–307.

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.

Frege, G. (1952). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of Gottlob Frege.* Blackwell, Oxford.

Gapp, K. (1995). Angle, distance, shape, and their relationship to projective relations. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society, CogSci-95*.

Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02*.

Gardent, C., Manuélian, H., Striegnitz, K., and Amoia, M. (2004). Generating definite descriptions: Non-incrementality, inference, and data. In Pechman, T. and Habel, C., editors, *Multidisciplinary Approaches to Language Production.* Mouton de Gruyter.

Garrod, S. C. and Sanford, A. J. (1982). The mental representation of discourse in a focused memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 35(1):21–41.

Gatt, A. (2006a). Generating collective spatial references. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society, CogSci-06.*

Gatt, A. (2006b). Structuring knowledge for reference generation: A clustering algorithm. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-06.*

Gluck, M. and Corter, J. (1985). Information, uncertainty and the utility of categories. In *Proceedings of the 7th Annual Conference of the Cognitive Science Society.*

Glucksberg, S., Krauss, R., and Higgins, E. T. (1975). The development of referential communication skills. In Horowitz, F. D., editor, *Review of Child Development Research*, volume IV. University of Chicago Press., Chicago & London.

Goodman, B. (1986). Reference identification and reference identification failures. *Computational Linguistics*, 12(4):273–305.

Gordon, P., Kendrick, R., Ledoux, K., and Yang, C. (1999). Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, 14:353–379.

Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics: Speech Acts.*, volume III. Academic Press.

Grice, P. (1957). Meaning. *The Philosophical Review*, 66:377–388.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Guhe, M., Habel, C., and Tschander, L. (2004). Incremental generation of interconnected preverbal messages. In Pechmann, T. and ., C. H., editors, *Current research on language production in Germany.*, Trends in Linguistics, Studies and Monographs 157. DeGruyter, Berlin.

Gupta, S. and Stent, A. J. (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Corpora in NLG, Birmingham, UK.*

Hajiĉova, E., Kubon, P., and Kubon, V. (1990). Hierarchy of salience and discourse analysis and production. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING-90.*

Hajiĉova, E. and Sgall, P. (2001). Topic-focus and salience. In *Proeedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL-01.*

Hajiĉova, E. and Vrbová, J. (1982). On the role of the hierarchy of activation in the process of natural language understanding. In *Proceedings of the 9th International Conference on Computational Linguistics, COLING-82.*

Hanna, J. E. and Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, 28(1):105–115.

Hanna, J. E., Tanenhaus, M. K., and Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49:43–61.

Heeman, P. and Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

Hermann, T. and Deutsch, W. (1976). *Psychologie der Objektbenennung.* Huber, Bern.

Hielscher, M. and Müssler, J. (1990). Anaphoric resolution of singular and plural pronouns: The reference to persons being introduced by different coordinating structures. *Journal of Semantics*, 7:347–364.

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-97.*, pages 206–213, Madrid.

Horacek, H. (2003). A best-first search algorithm for generating referring expressions. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL-03*.

Horacek, H. (2004). On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation, INLG-04*.

Horacek, H. (2005). Generation of referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation, ENLG-05*.

Itti, L. (2005). Models of bottom-up attention and saliency. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *Neurobiology of Attention*. Elsevier, San Diego, Ca.

Jackendoff, R. (1991). *Semantic structures.* MIT Press, Cambridge, Ma.

Jescheniak, J. D., Hantsch, A., and Schriefers, H. (2005). Context effects on lexical choice and lexical activation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31(5):905–920.

Johnson-Laird, P. (1983). *Mental Models.* Cambridge University Press.

Jordan, P. and Walker, M. (2000). Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.*

Jordan, P. W. (2000a). Can nominal expressions achieve multiple goals? In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Jordan, P. W. (2000b). Influences on attribute selection in redescriptions: A corpus study. In *Proceedings of the Cognitive Science Conference*.

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, Ca.

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.

Kaup, B., Kelter, S., and Habel, C. (2002). Representing referents of plural expressions and resolving plural anaphors. *Language and Cognitive Processes*, 17(4):405–450.

Kay, M. (1996). Chart generation. In *Proceedings of the 34th annual meeting of the Association* *for Computational Linguistics, ACL-96.*

Kelleher, J., Kruijff, G.-J., and Costello, F. (2006). Proximity in context: An empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL-COLING-06.*

Kelleher, J. D. and Kruijff, G.-J. (2006). 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.*

Keller, F. (2003). A psychophysical law for linguistic judgments. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society, CogSci-03.*

Kempen, G. and Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, 11:201–258.

Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, 24:253–270.

Keysar, B., Barr, D. J., Balin, J. A., and Baruner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89:25–41.

Kibble, R. (1999). Cb or not cb? Centering Theory applied to NLG. In *Proceedings of the ACL-99 Workshop on Discourse and Reference Structure.*

Kibble, R. and Power, R. (2000). An integrated framework for text planning and pronominalisation. In *Proceedgins of the 1st International Conference on Natural Language Generation.*

Kilgarriff, A. (2003). Thesauruses for natural language processing. In *Proceedings of the Conference on Natural Language Processing and Knowledge Engineering, NLP/KE-03.*

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. In *Proceedings of the 11th International Congress of the European Association for Lexicography, EURALEX-04.*

Kilgarriff, A. and Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL-01 Collocations Workshop.*

Koh, S. and Clifton, C. (2002). Resolution of the antecedent of a plural pronoun: Ontological categories and predicate symmetry. *Journal of Memory and Language*, 46:830–844.

Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation.* Stanford: CSLI.

Krahmer, E. and van der Sluis, I. (2003). A new model for generating multimodal referring expressions. In *Proceedings of the 9th European Workshop on Natural Language Generation, ENLG-03.*

Krahmer, E., van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Krauss, R. and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114.

Krauss, R. and Weinheimer, S. (1966). Concurrent feedback, confirmation and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343–346.

Krauss, R. and Weinheimer, S. (1967). Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behaviour*, 6:359–363.

Kripke, S. A. (1980). *Naming and Necessity.* Harvard University Press, Cambridge, Ma.

Kronfeld, A. (1989). Conversationally relevant descriptions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

Lakoff, G. and Johnson, M. (1981). *Metaphors we live by.* University of Chicago Press, Chicago, Il.

Lang, E. (1984). *The semantics of coordination.* John Benjamins, Amsterdam.

Langkilde, I. (2000). Forest-based statistical language generation. In *Proceedings of the 1st Meeting of the North Americal Chapter of the Association for Computational Linguistics, NAACL-00.*

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05.*

Lapata, M., McDonald, S., and Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, EACL-99.*

Lesk, M. (1986). Automatic sense disambiguation using machine-readable dictionaries. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC-86.*

Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Science*, 3:223–232.

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

Levelt, W. M. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences*, 22(1):1–37.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL-98.*

Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning.*

Lloyd, P. and Banham, L. (1997). Does drawing attention to the referent constrain the way in which children construct verbal messages? *Journal of Psycholinguistic Research*, 26(5):509–518.

Mangold, R. and Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, 7(3–4):181–191.

McCoy, K. F. and Strube, M. (1999a). Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of the ACL-99 Workshop on the Relation of Discourse/Dialogue Structure and Reference.*

McCoy, K. F. and Strube, M. (1999b). Taking time to structure discourse: Pronoun generation beyond accessibility. In McCoy and Strube (1999a).

McDonald, D. D. (1987). No better, but no worse, than people. In *Proceedings of the 3rd Workshop on Theoretical Issues in Natural Language Processing, TINLAP-87.*

McDonald, D. D. (2000). Natural language generation. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing.* Marcel Dekker, New York.

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):1–34.

Meteer, M. W. (1991). Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304.

Metzing, C. 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.

Meyer, A., Slederink, A., and Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66:B25–B33.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture-word interference experiments. *Journal of Memory and Language*, 35:477–496.

Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Minnen, G., J. Carroll, J., and Pearce, D. (2001). Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Moxey, L., Sanford, A. J., Sturt, P., and Morrow, L. I. (2004). Constraints on the formation of plural reference objects: The influence of role, conjunction and type of description. *Journal of Memory and Language*, 51:346–364.

Murphy, G. (1984). Establishing and accessing referents in discourse. *Memory and Cognition*, 12:489–497.

Murphy, G. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language*, 29(3):259–288.

Murphy, G. L. (2002). *The big book of concepts.* MIT Press, Cambridge, Ma.

Naor-Raz, G., Tarr, M., and Kersten, D. (2003). Is color an intrinsic property of object representation? *Perception*, 32:667–680.

Novak, H.-J. (1988). Generating referring phrases in a dynamic environment. In Zock, M. and Sabah, G., editors, *Advances in Natural Language Generation*, volume II. Pinter, USA.

Oberlander, J. (1998). Do the right thing ... but expect the unexpected. *Computational Linguistics*, 24(3):501–507.

O'Donnell, M., Cheng, H., and Hitzeman, J. (1998). Integrating referring and informing in np planning. In *Proceedings of the COLING-ACL Workshop on the Computational Treatment of Nominals*.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77:257–273.

Paiva, D. S. (1998). A survey of applied natural language generation systems. Technical Report ITRI-98-03, Information Technology Research Institute, University of Brighton.

Papineni, S., Roukos, T., Ward, W., and Zhu., W. (2002). Bleu: a. method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02.*

Paraboni, I., Masthoff, J., and van Deemter, K. (2006). Overspecified reference in hierarchical domains: Measuring the benefits for readers. In *Proceedings of the 4th International Conference* *on Natural Language Generation, INLG-06.*

Passonneau, R. J. (1995). Integrating Gricean and attentional constraints. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95).*

Passonneau, R. J. (1997). Interaction of discourse structure with explicitness of anaphoric noun phrases. In Walker, M., Joshi, A. K., and Prince, E., editors, *Centering in Discourse.* Oxford University Press, Oxford.

Pechmann, T. (1983). Accentuation and redundancy in children's and adults' referential communication. In Bouma, H. and Bouwhuis, D., editors, *Attention and Performance: Control of Language Processes*, volume X. Hillsdale: Erlbaum.

Pechmann, T. (1984). Accentuation and redundancy in children's and adults' referential communication. In Bouma, H. and Bouwhuis, D. G., editors, *Attention and Performance*, volume 10. Lawrence Erlbaum, Hillsdale, NJ.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

Pederson, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity — measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI-04.*

Pickering, M. and Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27:169–226.

Pollack, M. E. (1991). Overloading intentions for efficient practical reasoning. *Noûs*, 25(4):513–536.

Preparata, F. P. and Shamos, M. A. (1985). *Computational Geometry.* Springer.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36:1389–1401.

Pustejovsky, J. (1995). *The Generative Lexicon.* MIT Press, Cambridge, Ma.

Quine, W. V. A. (1953). *From a Logical Point of View.* Harvard University Press, Cambridge, Ma.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proc. 28th Annual Meeting of the Association for Computational Linguistics.*

Reiter, E. (1994). Has a consensus NL generation architecture appeared? And is it psycholinguistically plausible? In *Proceedings of the 7th. International Workshop on Natural Language generation, IWNLG-94.*

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Journal of Natural-Language Engineering*, 3:57–58.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press, Cambridge, UK.

Reiter, E., Robertson, R., and Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

Reiter, E. and Sripada, S. (2002a). Human variation and lexical choice. *Computational Linguistics*, 28:545–553.

Reiter, E. and Sripada, S. (2002b). Should corpora texts be gold standards for nlg? In *Proceedings of the 2nd International Conference on Natural Language Generation, INLG-02.*

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In

Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI-1995.

Rock, I. (1983). *The Logic of Perception.* MIT Press, Cambridge, Ma.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42:107–142.

Roelofs, A. (2000). Weaver++ and other computational models of lemma retrieval and word form encoding. In Wheeldon, L., editor, *Aspects of Language Production.* Psychology Press, UK.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4(3):328–350.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rosenberg, S. and Markham, B. (1971). Choice behaviour in a referentially ambiguous task. *Journal of Personality and Social Psychology*, 17:99–105.

Roy, D. (2002). Learning words and syntax for a visual description task. *Computer Speech and Language*, 16(3):353–385.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in the action and perception. *Artificial Intelligence*, 167(1–2):170–205.

Russell, B. (1905). On denoting. *Mind*, 14:479–493.

Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach.* Prentice Hall, New York, 3 edition.

Sanford, A. and Lockhart, F. (1990). Description types and method of conjoining as factors influencing plural anaphora: A continuation study of focus. *Journal of Semantics*, 7:365–378.

Sanford, A. and Moxey, L. (1995). Notes on plural reference and the scenario-mapping principle in comprehension. In C.Habel and G.Rickheit, editors, *Focus and cohesion in discourse.* de Gruyter, Berlin.

Sanford, A. J. and Garrod, S. C. (1981). *Understanding written language.* John Wiley and Sons, Chichester.

Sanford, A. J. and Moxey, L. M. (1999). What are mental models made of? In Rickheit, G. and Habel, C., editors, *Mental models in discourse processing and reasoning.* Elsevier Science, UK.

Schriefers, H. and Pechmann, T. (1988). Incremental production of referential noun phrases by human speakers. In Zock, M. and Sabah, G., editors, *Advances in Natural Language Generation*, volume 1. Pinter, London.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Oxford University Press, Oxford.

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.

Shasha, D. and Zhang, K. (1990). Fast algorithms for unit cost editing distance between trees. *Journal of Algorithms*, 11:581–621.

Shieber, S. M. (1993). The problem of logical form equivalence. *Computational Linguistics*, 19(1):179–190.

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-04.*

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):174–215.

Sonnenschein, S. (1982). The effects of redundant communication: When less is more. *Child Development*, 53:717–729.

Sonnenschein, S. (1984). Why young listeners do not benefit from differentiating verbal redundancy. *Child Development*, 55(3):929–935.

Sonnenschein, S. and Whitehurst, G. J. (1984). Developing referential communication skills: The interaction of role switching and difference rule training. *Journal of Experimental Child Psychology*, 38:191–207.

Spivey, M., Tyler, M., Eberhard, K., and Tanenhaus, M. (2001). Linguistically mediated visual search. psychological science, 12, 282-286. *Psychological Ssicence*.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Stone, M. (2000). On identifying sets. In *Proceedings of the 1st International Conference on Natural Language Generation, INLG-00*.

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, M. and Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In *Proceedings of the International Workshop on Natural Language Generation, INLG-98*.

Strawson, P. F. (1950). On referring. *Mind*, 59:320–344.

Suppes, P. (1971). Young children's comprehension of logical connectives. *Journal of Experimental Child Psychology*, 21:304–317. [Reprinted in Suppes (1991)].

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. G. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., and Morris, R. K. (2000). Priming in sentence processing: Intralexical spreading activation, schemas and situation models. *Journal of Psycholinguistic Research*, 29(6):581–595.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

van Deemter, K. (2000). Generating vague descriptions. In *Proceedings of the First International Conference on Natural Language Generation, INLG-00*.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

van Deemter, K. (2004). Finetuning NLG 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, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

van Deemter, K. and Halldórsson, M. (2001). Logical form equivalence: The case of referring expressions generation. In *Proceedings of the European Workshop on Natural Language Generation, ENLG-01*.

van Deemter, K. and Krahmer, E. (2006). Graphs and booleans: On the generation of referring expressions. In Bunt, H. and Muskens, R., editors, *Computing Meaning*, volume III. Kluwer Academic Publishers, Dordrecht.

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.*

van der Sluis, I. (2005). *Multimodal Reference. Studies in Automatic Generation of Multimodal Referring Expressions.* PhD thesis, Tilburg University, Tilburg, The Netherlands.

van der Sluis, I., Gatt, A., and van Deemter, K. (2006). Manual for the tuna corpus: Referring expressions in two domains. Technical report, University of Aberdeen.

van der Sluis, I., Gatt, A., and van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. Submitted.

van der Sluis, I. and Krahmer, E. (2005). Towards the generation of overspecified multimodal referring expressions. In *Proceedings of the Symposium on Dialogue Modelling and Generation, 15th Annual Meeting of the Society for Text and Discourse, STD-05.*

van Vliet, S. (2002). Overspecified NPs marking conceptual shifts in narrative discourse. In Broekhuis, H. and Fikkert, P., editors, *Linguistics in the Netherlands 2002.* John Benjamins, Amsterdam.

Varges, S. (2004). Overgenerating referring expressions involving relations and booleans. In *Proceedings of the 3rd International Conference on Natural Language Generation, INLG-04.*

Varges, S. (2005a). Chart generation using production systems. In *Proceedings of the 10th European Workshop on Natural Language Generation, ENLG-05.*

Varges, S. (2005b). Spatial descriptions as referring expressions in the Map Task domain. In *Proceedings of the 10th European Workshop on Natural Language Generation, ENLG-05.*

Varges, S. and Mellish, C. (2001). Instance-based natural language generation. In *Proceedings of the 2nd Meeting of the North Americal Chapter of the Association for Computational Linguistics, NAACL-01.*

Viethen, J. and Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation, INLG-06.*

Vigliocco, G., Lauer, M., Damian, M., and Levelt, W. J. (2002). Semantic and syntactic forces in noun phrase production. *Journal of Experimental Psychology: Human Perception and Performance*, 28:46–58.

von Stutterheim, C., Mangold-Allwinn, R., Barattelli, S., Kohlmann, U., and Kölbing, H.-G. (1993). Reference to objects in text production. *Belgian Journal of Linguistics*, 8:99–125.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In Ellis, W., editor, *A Source Book of Gestalt Psychology.* Routledge & Kegan Paul, London.

Whitehurst, G. and Sonnenschein, S. (1978). The development of communication: Attribute variation leads to contrast failure. *Journal of Experimental Child Psychology*, 25:490–504.

Whitehurst, G. J. (1976). The development of communication: Changes with age and modeling. *Child Development*, 47(473–482).

Wittgenstein, L. (1953 [2001]). *Philosophical Investigations.* Blackwell, Oxford.