Categories, Concepts and Co-occurrence:

Modelling Categorisation Effects with LSA

Louise M. Connell

M.Sc. in Cognitive Science and Natural Language
Division of Informatics
University of Edinburgh
2000

Abstract

Categorisation is a process that has been commonly tied to concepts and conceptualisation, as well as intimately linked with similarity. The underlying assumption in much of the literature is that the empirical evidence for categorisation effects has been the result of conceptual processing. This thesis questions this assumption by simulating such effects by the use of a co-occurrence model of language (LSA). Despite being a statistical tool based on simple word co-occurrence, LSA successfully simulates subject data relating to categorisation tasks, typicality effects and the effects of context on categories. The model is also used to successfully predict subject judgements of typicality in the presence of context. By virtue of these successes, this thesis argues that the nature of the representations used in conceptual thought in such categorisation tasks is open to debate and that another, context-based explanation for categorisation may exist.

Acknowledgements

I would firstly like to thank my supervisor, Dr. Michael Ramscar, for providing a seemingly endless supply of inspiration, ideas and feedback regarding the research reported in this thesis. I would also like to thank Dan Yarlett for his feedback and encouragement on the literature review portion of this tome.

Also, many thanks to Dermot Lynott for proof-reading, frequent advice and support during the course of this work, and generally keeping me grounded in sanity for the entire M.Sc.

To all the subjects who took part in my experiments, I give my thanks, especially to those whom I press-ganged into participation during their holidays.

My family also deserves my thanks for general support down through the years, and for encouraging me to go in my own direction even though it looked like I'd never have a real job. That has always been my aim.

Finally, more general gratitude goes to Barr, the makers of Irn Bru, without which I would have functioned far less efficiently in the latter stages of this thesis. And to Marie-Jeanne; it's been a good year.

Contents

CHAPTER 1 INTRODUCTION 1

BACKGROUND AND MOTIVATIONS 1
OVERVIEW OF THESIS 2

CHAPTER 2 CATEGORISATION, CONCEPTS, SIMILARITY 4

INTRODUCTION 4
THE CLASSICAL VIEW 6
Arguments for the Classical View 7
Flaws of the Classical View 9
THE FAMILY RESEMBLANCE VIEW 11
Arguments for the Family Resemblance View 12
Flaws of the Family Resemblance View 13
HYBRIDS OF CLASSICAL AND FAMILY RESEMBLANCE VIEWS 18
THE EXEMPLAR VIEW 19
Arguments for the Exemplar View 20
Flaws of the Exemplar View 22
HYBRIDS OF FAMILY RESEMBLANCE AND EXEMPLAR VIEWS 24
THE THEORY VIEW 24
Arguments for the Theory View 26
Flaws of the Theory View 28
WHERE SIMILARITY COMES IN 29
Similarity and Categorisation – not the same thing 29
Similarity and Categorisation – separate but intertwined 31

CHAPTER 3 CO-OCCURRENCE MODELS 35

INTRODUCTION 35
CO-OCCURRENCE TECHNIQUES 35
CATEGORISATION AND CO-OCCURRENCE 37
Revising Theories of Concepts 39

CHAPTER 4 MODELLING CATEGORISATION EFFECTS 41

INTRODUCTION 41
SIMULATION 1 – DEMONSTRATING BASIC CATEGORISATION 41
Method 42
Results 42
Discussion 44
SIMULATION 2 – DEMONSTRATING TYPICALITY IN CATEGORISATION 46
Method 47
Results 48
Discussion 50

SIMULATION 3 – DEMONSTRATING TYPICALITY IN WELL-DEFINED CATEGORIES 55
Method 56
Results 56
Discussion 57
SIMULATION 4 – DEMONSTRATING BASIC CONTEXT EFFECTS 58
Method 58
Results 60
Discussion 60
EXPERIMENT 1 – TYPICALITY IN CONTEXT 62
LSA Method 63
Subject Method 65
Subjects 65
Materials 65
Procedure 65
Results 66
Discussion 67
EXPERIMENT 2 – CONTEXTUAL TYPICALITY VS. CANONICAL TYPICALITY 72
Method 72
Subjects 72
Materials 73
Procedure 73
Results 73
Discussion 75

CHAPTER 5 GENERAL DISCUSSION 78

INTRODUCTION 78
CAVEATS OF CO-OCCURRENCE MODELS 78
A CONTEXT-BASED THEORY OF CATEGORISATION? 81
The Basis and Strengths 81
Natural versus Artifact Kinds 84
Where Concepts come in 86
CONCLUSIONS 87

BIBLIOGRAPHY 89

APPENDIX A 95

APPENDIX B 96

APPENDIX C 101

APPENDIX D 103

APPENDIX E 105

APPENDIX F 106

List of Figures

Figure 2.1: a small semantic net of the *animal* taxonomy..8

Figure 2.2: linear separability in AND, OR and inseparability in XOR.....................16

Figure 2.3: a Bedlington Terrier – note resemblance to a lamb 32

Figure 4.1: graph of city / country categorisation..43

Figure 4.2: graph of animal / body categorisation..43

Figure 4.3: graph of Rosch (1973) and LSA scores for category *sport*.......................52

Figure 4.4: graph of Rosch (1973) and LSA scores for category *crime*......................52

Figure 4.5: graph of Rosch (1973), Armstrong, Gleitman & Gleitman (1983) and LSA scores for category *vehicle* 52

Figure 4.6: graph of scaled LSA scores and subject ratings for *beverage* in context (agreement on 5/10 items – only context set that failed to cluster significantly) 69

Figure 4.7: graph of scaled LSA scores and subject ratings for *bird* in context (agreement on 9/10 items – clusters significantly at *p*<0.05 without *peacock*) 69

Figure 4.8: graph of subject scores for canonical and contextual typicality of s*port* (note visible differences between canonical and contextual typicality) 76

Figure 4.9: graph of subject scores for canonical and contextual typicality of *bird* (note lack of correlation between canonical and contextual typicality)................76

List of Tables

Table 4.1: Rank correlation coefficients *rho* (with levels of significance *p*) between LSA and subject scores and between Set A / Set B where applicable 49

Table 4.2: BNC frequencies and subject typicality .... scores for “well-defined” categories *odd number* and *even number* 57

Table 4.3: Referent items for two example context sentences 59

Table 4.4: LSA scores for referent items against context sentences 1 and 2 (from Table 4.3), with reaction times (RT) 60

Table 4.5: Sample context set with appropriate / inappropriate clusters and LSA scores 64

Table 4.6: Context sentences for LSA, subject scores, adjusted sets giving Wilcoxon’s *W* and significance of difference between clusters 67

Table 4.7: Rank correlations *rho* with significance *p* for contextual (Experiment 1) and canonical (Experiment 2) typicality scores 74

Chapter 1   Introduction

> "What, exactly is meant by the word 'category', whether in Aristotle or in Kant and Hegel, I must confess that I have never been able to understand."
>
> – Bertrand Russell: History of Western Philosophy

Background and Motivations

The human cognitive faculty of categorisation has a long history of research in psychology, with theories of how it operates moving through successive levels of sophistication according to emerging empirical data. Commonly linked to the process of categorisation is the representational question of concepts, with even researchers from opposing views assuming that a theory of one provides for the other (Armstrong, Gleitman & Gleitman, 1983; Keil, 1987; Lakoff, 1987a, 1987b). This is an assumption that has sustained confidence in the empirical methods used in the categorisation literature, which have presupposed that their task demands are conceptual in nature.

Related work on the connection between similarity and categorisation (Medin, 1989; Medin & Wattenmaker, 1987; Hampton, 1987b; Tversky, 1977) has illustrated the extent to which the two aspects of cognition are mutually reliant, and Hahn and Chater, (1997) have called for a measure of constrained similarity on which to found a cohesive explanation of conceptualisation. Recent work by Ramscar and Yarlett (2000; also Yarlett & Ramscar, 2000) has used a metric of semantic similarity based on co-occurrence techniques to simulate retrieval from long-term memory for analogical processes. This suggests that a model of co-occurrence could provide the necessary similarity constraint on which categorisation could be simulated.

However, co-occurrence models such as LSA (Landauer & Dumais, 1997) or HAL (Burgess & Lund, 1997) are essentially statistical tools that work on the premise that similar words are used in similar contexts. They exploit the frequency count of surrounding words for each lexeme to build a representation of meaning that is purely grounded within the language. Any human categorisation data that can be modelled by a mathematical algorithm based on word co-occurrence is therefore open to the question of whether it is a product of conceptual thought. The objective of this thesis was to examine the extent to which co-occurrence techniques could model this human categorisation data.

Overview of Thesis

The next chapter takes the form of a review of the categorisation and concepts literature, and discusses the relationship between similarity and categorisation. The following chapter then gives a brief overview of co-occurrence models, discussing the reason for their suitability in modelling categorisation effects.

Chapter 4 moves onto the simulation and experimental work, where the co-occurrence model LSA is shown to successfully simulate subject data relating to categorisation tasks, typicality effects (Rosch, 1973; Armstrong, Gleitman & Gleitman, 1983; Malt & Smith, 1984), and the effects of context on categories (Roth & Shoben, 1983). Typicality data on "well-defined" categories from Armstrong, Gleitman & Gleitman (1983) is also shown to be due to word frequency. The first empirical experiment shows that LSA can be used to predict subject typicality ratings for items with the presence of context, where each item is either appropriate or inappropriate in the given context sentence. The second experiment then confirms that these subject contextual typicality ratings are significantly different from canonical typicality (where no context is given).

The final chapter considers these findings, noting the limitations of the LSA model in its current form. A context-based theory of categorisation is offered, by way of explaining how a statistical data analysis tool can be used to model empirical work hitherto considered as conceptually grounded. The general conclusion is that the data modelled here – and hence much of the categorisation literature – is subject to the question of whether it has been based on too simple a view of conceptual thought.

Chapter 2 Categorisation, Concepts, Similarity

Introduction

Categorisation is sorting. This may be its most concise and simplistic definition, yet at least it is one that manages to encompass the diverse theories of categorisation without contradicting any of their basic tenets. A more complex definition mentioning similarity or shared features will not be acceptable to the same extent, simply because there is so little agreement between theories on the operation of the essential human ability to generalise and classify. Funes the Memorious, the creation of J. L. Borges (1964), was a man afflicted by the inability to generalise what he saw. Lacking even the capacity to associate instances of the same object separated by time, the unfortunate Funes was utterly incapable of generalising different objects and grouping them into categories. Without generalisation, he was without conceptualisation, and ultimately without normal human function.

The terms *categories* and *concepts* are often used interchangeably in the literature (e.g. Armstrong, Gleitman & Gleitman, 1983; Keil, 1987) and in different senses depending on whether the field in question is psychology, linguistics or philosophy. In this review, the author follows a common assumption (Medin & Smith, 1984; Komatsu, 1992) that categories are classes, and that concepts are their mental representations. An instance is a specific example of a category member. This literature review looks at four principal theories of categorisation and concepts: the Classical (definitional), Family Resemblance (probabilistic), Exemplar (instance), and

Theory (explanation) theories, as well as their hybrids. All four theories are also discussed in relation to similarity.

The first attempts to analyse the basic cognitive faculty of categorisation came in the works of Plato and Aristotle, where the latter put forward what became known as the classical view, namely that categories are differentiated by defining attributes. This theory of categories reigned predominantly for twenty centuries, being further refined by research in the 20th century (such as Katz & Fodor, 1963; Katz, 1972) to state that the representation of a concept consists of a set of necessary and sufficient features. In the classical view, the concept *cat* consists of information about the necessary and sufficient attributes of cats – [fur, four-legs, tail, whiskers, …etc.].

In the 1970's, the classical view fell into disfavour with the ascendance of the family resemblance theory (Rosch & Mervis, 1975). This term was borrowed from Wittgenstein's (1953) analysis of concepts and categories –however, henceforth the author refers to 'family resemblance' in the Roschian rather than Wittgensteinian sense. This view is based around trying to account for typicality effects – the phenomenon of category gradedness where subjects were found to judge some instances as highly typical (prototypical) of a category and others less so. In a family resemblance view, the concept *cat* consists of an average abstracted summary of instances of cats encountered, giving certain attributes more weight than others.

An alternative account of categorisation that arose at a similar time was the exemplar view (Medin & Schaffer, 1978). It too provides an account for typicality effects, and differs from the family resemblance theory in terms of representation. With the exemplar view, the concept *cat* consists of past instances of cats previously encountered, rather than a summarised abstraction.

The 1980's saw the growing popularity of a different account of categorisation and concepts – the theory theory (Murphy & Medin, 1985; Lakoff, 1987a, 1987b), so called because it focuses on the information people have about relations between concepts and attributes, and offered explanations about why certain categories cohere. In the theory view, the concept *cat* is made up of information about instances of cats, their attributes, their interactions with the rest of the world, and the (often causal) relationships that exist between all these.

Reviews of most theories of concepts can be found in Medin (1989) and Komatsu (1992), the latter of which also mentions some of the hybrids that exist between many of these main four theories. These hybrids will be discussed in more detail later.

The Classical View

The idea that a category may be defined by a set of necessary and sufficient characteristics provided an account of concepts compelling enough to last 2,000 years. First to posit this view was Plato, but it was his student, Aristotle, that the theory is generally attributed to. Where Aristotle differed from his predecessors, most notably Plato, was in the source of categorical knowledge. Plato was an essentialist, meaning that, like Aristotle, he not only believed that "things" are defined by a set of necessary features, but that this essence existed separately from the "things" themselves. Plato posited a realm of ideals, from which the objects in the world are imperfect reflections, and it is from these ideals that the objects take their essence. Aristotle's alternative was that it is not necessary to hold that the essences of things exist in some ideal realm: he said that the essences are simply part of our knowledge of the world. We know what makes something a *cat*, what its defining attributes are, and we can deduct this knowledge from the examples all around us. Aristotle's theory of categories was regarded as a body of unquestionable truth for centuries. Later refinements took many forms, such as Katz and Fodor (1964) who attempted to model the semantics of natural language in terms of feature sets and restrictions.

Arguments for the Classical View

Many points exist in the classical view's favour. Firstly, classical concepts are very economic in their representation, having to store only a set of individually necessary and collectively sufficient features for each concept. To cite a favourite example of the literature, the concept *bachelor* may be defined by the attributes [human, male, adult, never-married]. It is a tidy definition that allows us to identify bachelors while sticking to the principle of cognitive parsimony. When considering the large number of categories into which we can place a familiar item (a *bachelor* is also a *man*, a *human*, a *mammal*, an *organic life form* etc.), it becomes clear how vast is the number of concepts we would potentially require over a lifetime. A small, neat representation for each concept becomes desirable.

Also important for the classical view is its strong account of category coherence – how it explains why the members of a category gel together and are separated from non-members. Any item that possesses the attributes [human, male, adult, never-married] can be clearly categorised as a *bachelor*, and it is separable from any item that does not share these exact attributes. This classical account of coherence is also connected to that of semantic networks, the hierarchical taxonomy of Collins and Quillian (1969). The distances between nodes in a semantic net like that of Figure 2.1 were found to correlate with subjects' response times in sentence verification tasks such as 'Does a bird have feathers' (short distance, short time) and 'Is a bird an animal' (longer distance, longer time). This was taken as validation of the classical view, which was also thought to possess a certain intuitive appeal. People prefer to think of categories as being definable, even if they cannot provide these definitions (McNamara & Sternberg, 1983).

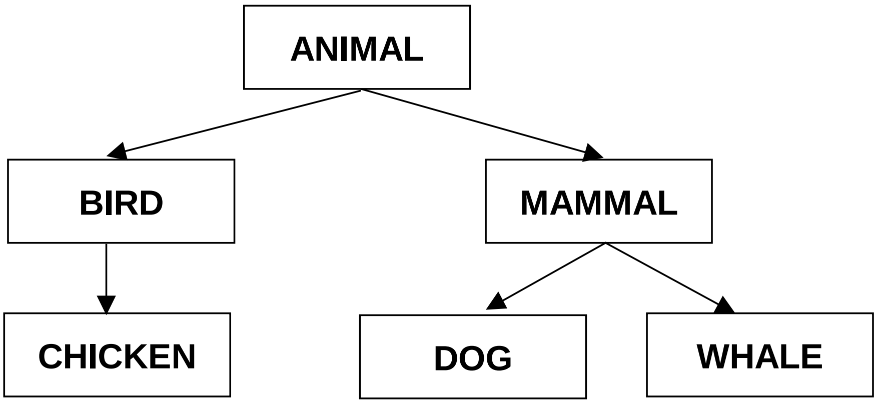

Figure 2.1: a small semantic net of the *animal* taxonomy

Flaws of the Classical View

However, during the 1970's empirical evidence emerged that could not be accounted for by the classical theory. Foremost was the discovery of typicality effects (Rosch, 1973, 1975a, 1975b), where subjects were found to judge some category members as more (proto)typical than others. Rosch (1973) gave subjects a category name (such as *fruit*) with a list of members (such as *apple*, *fig*, *olive*, *plum*, *pineapple*, *strawberry*), and asked subjects to rate on a 7-point scale how good an example each member was of its category. The results showed a clear trend of category gradedness – apples are consistently judged a typical *fruit*, while olives are atypical. This finding is in direct contrast with the classical notion of discrete categories, which ascribes membership an all-or-nothing status. Further evidence underlined the pervasiveness of typicality (or 'goodness of example'), and its ability to predict a variety of results. Level of typicality was found to predict reaction times in sentence verification tasks (Rosch, 1973; Rosch & Mervis, 1975; McCloskey & Glucksberg, 1979) and order of item output when subjects are asked to name members of a category (Barsalou & Sewell, 1985). Typicality has even been successfully applied to emotion terms (Fehr, 1988) and artistic style (Hartley & Homa, 1981). To return to our earlier classical concept of *bachelor*, it is possible to find typical and atypical members of this apparently definable category. A typical *bachelor* instance may be Humphrey Bogart's character of Rick Blaine in *Casablanca*. However, even though he may fulfil the necessary and sufficient conditions [human, male, adult, never-married], the pope would be considered highly atypical.

Related to its assumption of discrete categories, the classical view also assumes that membership is clear-cut. However, when McCloskey & Glucksberg (1978) asked subjects to categorise certain familiar objects (such as *rug*, *clock* or *radio* as *furniture*), they found considerable disagreement between subjects and even within subjects. Across a one-month period, some 22% of subjects' answers changed when asked to categorise the same items. Not only do people disagree about what category items belong to, but contradict themselves on different occasions. This implies that categories have fuzzy rather than clear-cut boundaries (see also Barsalou, 1989).

Also problematic for the classical theory is subjects' frequent inability to name necessary and sufficient conditions for categories (McNamara & Sternberg, 1983), and that when they can, results show substantial inter- and intra-subject disagreement (Rosch & Mervis, 1975; Komatsu, 1983). For example, subjects may list [made-of-wood] as a necessary property for *violin*, but as all violins are not made of wood, this is not a necessary condition and would lead to the exclusion of a number of valid members. Alternatively, subjects may list the attribute [unmarried] among those for *bachelor*, which would allow the admittance of a widowed or divorced man into the category. Thus [unmarried] is not a sufficient characteristic, and would lead to the inclusion of non-members.

Finally, the semantic networks of Collins and Quillian (1969) came under scrutiny. Smith, Shoben & Rips (1974) showed that in some cases, in comparing node distance and sentence verification times, the proportional relationship collapses. For example, subjects were quicker to answer yes to 'A chicken is an animal' than 'A chicken is a bird', despite the fact that the taxonomy flows chicken-bird-animal. This runs directly contrary to the Collins & Quillian findings. Hampton (1982) also found that the common assumption implicit in semantic nets – that taxonomies represent class inclusion hierarchies – does not always hold true. His subjects agreed that a *chair* was a type of *furniture,* and that a *carseat* was a type of *chair*. However, they were unwilling to allow a *carseat* into the category *furniture*. This category intransitivity negates the notion of complete inheritance from superordinate classes.

The Family Resemblance View

In order to address some of the classical view's problems, most notably typicality effects, Rosch & Mervis (1975) proposed the family resemblance theory of concepts. Citing Wittgenstein (1953), they argue that what essentially links members of a class is a family resemblance to each other. In practise, this means that instead of defining a category by a set of necessary and sufficient attributes, possessing only some of the attribute set is deemed sufficient. No attributes are deemed necessary, but some are more heavily weighted than others. Attribute weighting is a measure of salience, and is calculated based on the number of category members that share that attribute. For example, since most members of the category *bird* share the attribute [feathers], this would be highly salient and heavily weighted. Fewer members share the attribute [can-fly] and this attribute would have lower weighting, and so on. A family resemblance entity consists of a set of shared attributes that are weighted for salience, and basically embodies an abstracted summary, or average, of the instances previously encountered. Rosch (1978) claims that this entity does not constitute a concept – explicitly denying that family resemblance theory postulates anything about representation and is rather a theory of categorisation – but others (Armstrong, Gleitman & Gleitman, 1983; Smith & Medin, 1981) do indeed treat this entity as a family resemblance concept. When a new item is encountered, its attributes are weighted and summed. If the total is above a required threshold, the item has a high degree of family resemblance and belongs in the category, otherwise it is rejected.

Arguments for the Family Resemblance View

The intention of Rosch and Mervis was to formulate a theory of categorisation that holds typicality at its core. The family resemblance view revolves around this phenomenon – centrality of typicality – where degree of typicality is directly related to degree of family resemblance. An instance possessing attributes that are shared by most other members of the category will have a high weighting score, and thus a high degree of family resemblance. The higher the degree of family resemblance an instance has in the category, the more typical it will be. Rosch and Mervis's (1975) theory of categorisation not only explains the ubiquitousness of typicality effects, but the same framework also provides an account of fuzzy boundaries between categories. Noise or variability in the salience weights of any attribute would lead to fuzziness around category thresholds (Smith & Medin, 1981), creating a degree of ambiguity between members and non-members whose summed weights are near this threshold level. A *bat* may hover on the edge of the category threshold for *bird*, as it shares many of the category's attributes [wings, can-fly] etc. However, lacking a heavily weighted attribute such as [feathers] should be enough to exclude it. The use of a threshold also allows the family resemblance view to provide an account of category coherence. Any items belonging to a category gel with the other members by virtue of their summed weights scoring above the category threshold, which in turn separates them from non-members (the threshold in question is calculated according to the weights of previously categorised members). By considering an item against individual possible categories, the category intransitivity problem of the classical view (Hampton, 1982) does not arise. A *carseat* may score above the threshold for *chair*, and be categorised as such. However, *carseat* may fail to score above the threshold for *furniture*, even though *chair* itself does.

This family resemblance set of abstracted common attributes and their weights is also still quite economical in terms of representation, if not quite as parsimonious as the classical theory. Storing extra information about a concept is the price paid for its new flexibility. In this way, family resemblance theory implicitly provides a tidy explanation for subjects' difficulty naming necessary and sufficient conditions (Rosch & Mervis, 1975; McNamara & Sternberg, 1983; Komatsu, 1983) – there are none. Even the most heavily weighted attribute may not be shared by all category members – not all fruit is sweet.

Flaws of the Family Resemblance View

However, despite the family resemblance view's focus on typicality, the theory has been challenged for not always being able to adequately explain typicality effects. Using defined categories such as *odd number*, Armstrong, Gleitman and Gleitman (1983) showed that subjects still rated instances by membership gradedness. The number 7 was judged highly typical, while the number 57 was not. As the only required attribute for class membership was for the number to be odd (divisible by 2 with a remainder of 1), then all members of the category share this attribute alone and there are no other attributes on which to base salience weightings. Thus, this difference in typicality cannot be aligned with a difference in family resemblance as it can with categories that share many attributes.

Other research (Osherson & Smith, 1981; Medin & Shoben, 1983) also showed that typicality in combined concepts cannot be reliably predicted from the typicality of its constituents. A typical *pet* may be a *dog* and a typical *fish* may be a *salmon*, but this makes it difficult to explain why a typical *pet fish* may be a *goldfish*. In a more complex example for the concept *spoon*, subjects judge small spoons more typical than large ones, and metal spoons more typical than wooden ones. The family resemblance view would infer that the spoons with the greatest degree of family resemblance (and hence highest typicality) would be small metal ones, and those with the least family resemblance (lowest typicality) would be large wooden ones, with other combinations falling in between. In fact, what Medin and Shoben (1983) found was that large wooden spoons were considered the second most typical *spoon* type after small metal ones, not the least typical. Family resemblance theory cannot explain this.

Malt and Smith (1983) note similar findings, and explain the family resemblance view's predictive failure in terms of insensitivity to attribute correlation. Weighted attributes in a family resemblance set are considered to be independent of each other, but this is not necessarily true. Subjects know that in terms of *spoon* types, [large, wooden] are correlated attributes in a way that [small, wooden] are not. In other words, certain attributes occur with certain others with a degree of regularity. If an item has fins we know that it can probably swim, but this information is not captured by the family resemblance view. The assumption of independent attributes also limits the type of categories that the family resemblance view can cover. Medin and colleagues (Medin & Schaffer, 1978; Medin & Schwanenflugel 1981) showed that because of its addition of independent attributes, the family resemblance view could only distinguish between linearly separable categories. A simple example of a linearly inseparable category is XOR (exclusive-or), which for the conditions A and B is true if either A or B is true, but not both together. In a feature-space of these two conditions, a straight line cannot be drawn to separate the true from the false – i.e. they are not linearly separable, unlike simpler relations such as AND or OR (see Figure 2.2). Without considering the relations that exist between attributes, the family resemblance view cannot deal with all categories.

Until this point, all mention of typicality has been concerned with the typicality of an item in the general sense of its category – its canonical or context-free form. However, Roth and Shoben (1983) showed that the context a concept appears in affects the typicality of its instances. A typical *bird* in the context-free sense may be a *robin*, but if it appears in the context ‘The bird walked across the barnyard’, then *chicken* would instead be typical. Subject reaction times to sentence verification tasks are faster for the contextually appropriate item (*chicken*) than the normally typical, but contextually inappropriate item (*robin*). Roth and Shoben found that typicality, as determined in isolation, no longer plays an important role once context in introduced.

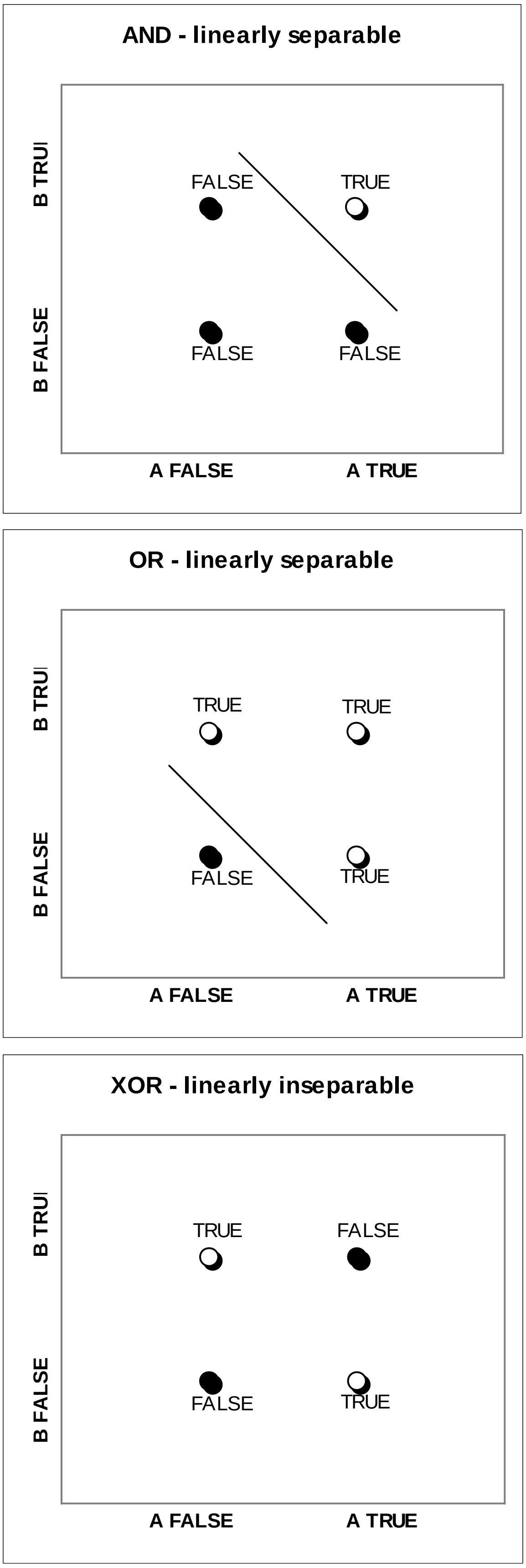

Figure 2.2: linear separability in AND, OR and inseparability in XOR

In some cases, they found it played no discernible role at all. This is incompatible with the family resemblance view's centrality of typicality. If context can skew, or even reverse typicality gradedness, then the notion of family resemblance by attribute weighting is negated.

Related to the issue of context, Barsalou (1983, 1987) introduced the idea of typicality in *ad hoc* or goal derived categories. In the same way that there are categories of *spoons* or *pet fish*, there are also ad hoc categories such as *things to take from one's home during a fire*, *things to put in a jumble sale*, or even something as arbitrary as *things that could fall on your head*. Barsalou found that subjects were as happy to produce typicality ratings for items in these ad hoc categories as they were for traditional taxonomic categories. Members of ad hoc categories often have few shared attributes at all – for example, *things to take from one's home during a fire* may include *children*, *photographs*, *pets* and *jewellery*, which at most may share the subjective attribute [valued]. Barsalou (1983) describes subjects' typicality ratings in these cases as the ability to work in novel contexts for which they do not have pre-stored graded structures. The lack of previous exposure to the category, as well as the shortage of shared attributes between members, means that the family resemblance view cannot explain the presence of typicality effects in ad hoc categories.

As well as the typicality problems with family resemblance, there are some issues regarding the acquisition of the set of attributes and their weightings. A family resemblance concept is a summarised abstraction, only storing shared attributes (i.e. not individualistic attributes or 'quirks'), and it is not clear how an attribute is determined to be potentially sharable as opposed to strictly individual. A quirk in one instance may appear again in future instances, thus deserving a weighted place in the attribute set. However, if the attribute from the original instance was not stored, then the future instances will also have the attribute dismissed as a quirk and the concept loses salient information. While perhaps not problematic in adults, this makes it difficult to determine how children would build a family resemblance concept.

Hybrids of Classical and Family Resemblance Views

The combination of advantages and problems in the family resemblance view is almost opposite to those of the classical view, which led some researchers to believe a hybrid of the two theories would be more successful. Subjects' tendency to believe in necessary and sufficient definition of categories (McNamara & Sternberg, 1983) and the demonstration of typicality effects in definable categories (such as *odd number*) (Armstrong, Gleitman & Gleitman, 1983) were taken to indicate the presence of both classical and family resemblance representations, each used for different tasks. Such hybrids (or dual-representational models) are usually based on the premise of using the family resemblance view for identifying members of the category, and the classical view to reason about the concept (Miller & Johnston-Laird, 1976; Osherson & Smith, 1981; Smith & Medin, 1981). Alternatively, the classical representation may be used for logical reasoning and the family resemblance representation for a kind of analogical reasoning (Rosch, 1983). However, these hybrids are generally prone to many of the same flaws as family resemblance, such as insensitivity to context and attribute correlation, inability to explain ad hoc categories, and inability to distinguish linearly inseparable categories.

The Exemplar View

In much of the literature (Smith & Medin, 1981; Lakoff, 1987a; Medin, 1989; Komatsu, 1992), there is disagreement over what constitutes an exemplar view. This is partly due to the minefield of contradictory uses of the terms *prototype theory* and *probabilistic theory*. Despite Rosch's (1978) statement that typicality effects do not constitute a theory of concepts or posit anything about the structure of concepts, they have often been interpreted as doing so. Prototype theory is the umbrella term sometimes used for any theory of categorisation that compares new items to a stored representation using some form of similarity, be it a summarised abstraction (family resemblance – Rosch & Mervis, 1975), a collection of instances (instance approach – Medin & Schaffer, 1978), a point in multidimensional psychological space (Generalised Context Model – Nosofsky, 1984, 1988) or instances at the centre of a radial structure (ideal cognitive models - Lakoff 1987a, 1987b). Additionally, family resemblance and instance approaches are frequently grouped under the other umbrella term of probabilistic theory. Exemplar theory itself has also been considered as an umbrella term for family resemblance and instance views together (Smith & Medin, 1981), as well as an equivalent name for the instance approach. The author will follow the example of Komatsu (1992) and Medin (1989) in considering the exemplar view *only* in terms of the instance approach (e.g. Medin & Schaffer, 1978; Nosofsky, 1984, 1988), distinguishing it from the family resemblance view already discussed. Lakoff's (1987a, 1987b) account the author considers a theory theory, and it is discussed under the relevant section.

The exemplar (instance) view differs from the family resemblance view in a number of ways. Firstly, the stored representation of a category is not an abstracted summary or shared features, but a set of previously encountered instances. The number of instances stored varies between accounts of the exemplar view from a core set to all previous instances. Abstraction does not take place at acquisition, but rather on retrieval for comparison with a novel item – i.e. the new item triggers the retrieval of instances, which are abstracted (averaged) on the fly. A subset (which may be the full number) of instances are retrieved, according to the novel item and affected by context, frequency, etc. Each instance may have full or partial information stored. Finally, shared attributes are combined multiplicatively, not additively as with the family resemblance view.

Arguments for the Exemplar View

Storing individual instances instead of an abstraction can explain why the accuracy of classification increases with category size (Busemeyer, Dewey & Medin, 1984). Also, by allowing the retrieval of a subset of these instances, the exemplar view is given much flexibility in explaining the typicality issues that the family resemblance view could not. Roth & Shoben (1983) showed that typicality effects were affected by the context in which the category appeared. The exemplar view holds that the novel item triggers retrieval of a certain subset of instances. If the context a novel item is embedded in influences the instances that are retrieved, then typicality effects based on that subset will differ from those of the canonical category. A context of walking across barnyards may cause the retrieval of a specific subset of *bird* instances

– perhaps excluding those who hop rather than walk (including the ubiquitous *robin*), and those incompatible with barnyard environments (such as the *penguin*) – with the most typical *bird* in the subset being *chicken* or *turkey*. With more than one instance retrieved at once, it is also possible to perform a simultaneous multiple comparison. This can explain subjects' sensitivity to attribute correlation (Malt and Smith 1983; Smith & Medin, 1981), because unlike with the family resemblance view, a subject is not limited to the concept representation abstracted on acquisition. An on-the-fly comparison of *spoon* instances will reveal that the attributes [large, wooden] appear to be correlated in a way that [small, wooden] are not, thus influencing typicality judgements for the category *spoon* in favour of large wooden ones. In the same way, exemplar theory can explain typicality in at least some forms of concept combination (Osherson & Smith, 1981; Medin & Shoben, 1988).

Assuming attributes are not independent also leads to their multiplicative, rather than additive, combination, which in turn leads to the exemplar view's explanation of linearly inseparable categories. The curved XOR boundary seen in Figure 2.2 cannot be made by linear (additive) combination, but is possible with quadratic (multiplicative) combination, so making the category separation possible (Medin & Schaffer, 1978; Medin & Schwanenflugel 1981; Nosofsky, 1986). The exemplar view may also be sensitive to goals in ad hoc categories (Barsalou, 1987, 1989). The goal *things to take from one's home during a fire* may cause the retrieval of instances from various other categories that are in some way considered irreplaceable, with typicality effects again emerging from the subset. Thus the exemplar view can deal with typicality in a larger range of category types than family resemblance.

Retrieval of category subsets also offers an explanation for subjects' tendency to believe in necessary and sufficient conditions, even if they cannot provide these definitions (McNamara & Sternberg, 1983). The exemplar view holds that on any given occasion, subjects may retrieve only a subset of their stored instances. Therefore, if they are inclined to regard this subset as exhaustive, false beliefs about clear category boundaries may emerge (Nickerson, 1981) – small finite sets are more likely to be encapsulated by a set of necessary and sufficient conditions that would not hold true with the rest of the category members. Storing individual instances rather than an abstraction has also led to challenges on the grounds that central tendencies are available long after information about specific instances has faded (Robbins, Barresi, Compton, Furst, Russo & Smith, 1978). Further challenges arose from the findings of Hayes-Roth and Hayes-Roth (1977) concerning the disassociation of classification and recognition. Here, subjects were more confident about their classification of prototypes than old instances, while still being more confident they had previously seen old instances rather than the prototypes. However, since the retrieved subset may contain partial instances (where not all information was stored), and may be biased by expectations, these results can be explained adequately by the on-the-fly abstractions of the exemplar view (Medin & Schaffer, 1978; Nosofsky, 1988).

Flaws of the Exemplar View

Although the exemplar view overcomes many of the problems of family resemblance, there are still some areas that cannot be so easily waved aside. The storage of many individual instances per category is not very economical in representational terms. Different versions of the exemplar view store instances differently – from every instance being partially stored (Reed, 1972) to most instances being stored at varying degrees of completeness (Medin & Schaffer, 1978) – but all of these require a lot more storage space than the concise classical or family resemblance entities.

Where the exemplar view suffers most is in failing to give an account of category coherence. A novel *bird* item is encountered, and so a subset of *bird* instances are retrieved and abstracted in some way and compared to the novel item. This comparison – using some form of similarity – is unconstrained. Goodman (1972) points out that if we say two objects are similar because they share many properties, then this quickly becomes meaningless as all entities have an infinite set of properties in common. A plum in my garden and a hydrogen atom in the sun's core both share the attribute of weighing less than 1kg (and 1.01kg, 1.001kg etc.) Likewise, all entities have an infinite number of properties *not* in common – a hydrogen atom weighs less than 1 gram (and 1.01g, 1.001g, etc.) while a plum weighs more. Without a constraint on what constitutes similarity, there is nothing in the exemplar view to explain why a *plum* belongs in the category *fruit* while a *hydrogen atom* does not.

The question of comparison also affects how the exemplar view may deal with definable categories such as *odd number* (Armstrong, Gleitman & Gleitman, 1983). If typicality effects were said to arise due to the retrieval of (a subset of) previously encountered *odd number*s, this would require the category *odd number* to be fully-formed. Yet for this category to cohere in the first place, why would *3* be judged more similar to *46827* than to *2* or *4*? By thinking in terms of similarity, then *3* could be grouped with *2* and *4* by virtue of being under *10*, *11*, *12*, etc. where *46827* cannot.

Again, without a constraint on what constitutes similarity, this makes it difficult for exemplar theory to explain how categories may form (be learned) at all.

Hybrids of Family Resemblance and Exemplar Views

It is this lack of category coherence that led to the hybridisation of exemplar and family resemblance representations. The schema approach (Rumelhart, 1980; Cohen & Murphy, 1984) is one such hybrid, and basically consists of storing a representation that captures the family resemblance abstraction of the concept along with information on the instances of the exemplar view. Also explicitly stored are logical and causal relationships between attributes. This offers the advantages in explaining typicality that the exemplar view does, but also allows the summarised abstraction to account for category coherence as it does with the family resemblance view. However, the schema approach is limited by its foundations in Artificial Intelligence (AI) frames, where slot-filling, default attribute values and inheritance in schematic networks are in many ways a reminder of the limitations of Collins & Quillian's (1969) semantic nets. The schema approach was mostly abandoned in favour of theory theories.

The Theory view

The theory, or explanation-based, view came into ascendance mainly due to the failure of highly-specified theories such as family resemblance (Rosch & Mervis, 1975), exemplar-based context models (Nosofsky, 1986, 1988), or schemata (Cohen & Murphy, 1984) etc. to fully capture the complexities of categorisation. There are many different flavours to theory theories, often only loosely aligned with each other (Murphy & Medin, 1985; Keil, 1986; Lakoff, 1987a, 1987b; Michalski, 1989; Wisniewski & Medin, 1994), but what they all have in common is this: they focus on the relationships that exist within and between concepts, and they focus on the host of 'background' knowledge that people employ when making any conceptual decision. A *bachelor* is not [human, male, adult, never-married]. Nor is a *bachelor* an abstraction or collection of items in the category. Rather, according to the theory view, a *bachelor* is a man that has never married but is of a marriageable age, of heterosexual disposition, and exists in a human society that both supports marriage and that provides enough eligible males and females for the practice to survive (Lakoff, 1987a). In other words, to understand what items belong in the category *bachelor*, we require a theory about social expectations and how different types of men fit these expectations.

What makes the 'concept' of the theory theory different from that of both the family resemblance and exemplar views is the question of concept stability. The entities of both the family resemblance and exemplar theories allow for some adjustment according to newly encountered items – a novel item may shift the weights of a family resemblance category, or may be added to the pool of exemplars. However, the inherent view is that these concepts are stable representations. Certain versions of the theory-based view (Barsalou, 1987; Michalski, 1989; Medin & Wattenmaker, 1987; Johnston-Laird, 1983) cast doubt on this assumption. By considering the various things that cause the structure of a category to change – including linguistic context (Roth & Shoben, 1983), decision-making about the nature of the category (Armstrong, Gleitman & Gleitman, 1983), points of view (Barsalou & Sewell, 1984), etc. – they suggest that categorisation is a result of constructing representations in working memory according to a particular context, drawing on some stored knowledge (theories). Thus, concepts are not stable representations, but rather are emergent entities formed on the fly from information in long-term memory (see also Ramscar & Hahn, 1998).

Arguments for the Theory View

The theory view is essentially less focussed on attributes and their combinations and more so on the relations that link concepts with each other and the rest of the world. Our theory about the concept *bird* tells us that a small bird (such as a wren) is more likely to sing than a large one (such as an ostrich) – this is a sensitivity to attribute correlation that is a by-product of our general theories about the *bird* category (Keil, 1989). This same relational information makes a concept sensitive to context, goals etc., so Roth and Shoben's (1983) *bird* walking across the barnyard will most typically be a chicken, because walking across barnyards forms part of our theory of what a *chicken* does. Our *penguin* theory does not link penguins with barnyards, so this would be a highly atypical choice of *bird* in this context. The large amount of relational knowledge inherent in any theory was also required to account for subjects' understanding of concept combinations (Medin and Shoben's 1988). The theory view grants an enormous degree of flexibility to concepts and what they may do, and also provides an account of category coherence – members group together because they each fit our theory of what the category is about. Explanatory relations link them.

As well as meeting the requirements set by earlier theories, the theory view offers some interesting observations. Medin, Wattenmaker & Hampson (1987) found that subjects in sorting tasks repeatedly failed to categorise cartoon creatures based on family resemblance, and only succeeded in classification when the interproperty relations – the underlying explanations for the category groups – were explained to them. Similarly, Wattenmaker, Dewey, Murphy & Medin (1986) showed that the ease with which subjects were able to learn different categories is affected more by the activation of possible underlying interproperty relations – explanations – than linear separability. These findings underline the importance of relational and extra 'background' knowledge to subjects' ability to categorise, something compatible with the theory view's tenets.

Returning full-circle through theories of categorisation to Plato's work, reviews of the theory approach (Medin, 1989; Komatsu, 1992; Hampton, 1997a) often align it with essentialism, although psychological rather than metaphysical. Metaphysical essentialism holds that things have essences that make them what they are – an *ostrich* has a 'bird' essence, a *rock* has a 'rock' essence – like Plato's theory that objects had an ideal form. Psychological essentialism (Medin & Ortony, 1989), on the other hand, does not hold that these internal essences actually exist, but just that people believe that they do. This is a useful standpoint (Komatsu, 1992) for explaining subjects' tendency to believe in necessary and sufficient conditions, even if they cannot provide these definitions (McNamara & Sternberg, 1983). If a person's representation of an object include the belief that the object belongs in a particular category by virtue of possessing an ‘essence’, then that essence would embody a necessary and sufficient condition.

Flaws of the Theory View

The principle flaw of the theory view is the lack of agreement between its practitioners. Lakoff (1987a, 1987b) calls his idealised cognitive model a prototype approach, Johnston-Laird’s (1980, 1983) mental models are described as being both schemata and theory-based by Komatsu (1992), etc. For this reason, most particular flaws of one theory-based approach may not be true of another. Their common flaw is a lack of specificity. Unconstrained relational links are subject to the same problems as unconstrained similarity (Goodman, 1972) – they can be of infinite number. For this reason, the theory view suffers at the very least from poor economy in representation. However, if relational information is truly unconstrained, then a single concept may incorporate all the information available to the person at the time through explanatory links to other concepts, equivalent to the frame problems in AI. The repercussion of this is that the line between background and concept-specific information is blurred, so every new piece of information a person acquires will have an unpredictable ‘ripple’ effect on the entire conceptual spectrum. The theory view must constrain its relational links in some way or subject the cognitive load to exponential increase. Among flavours of the theory view, there is no commonly agreed or fully adequate specification for this constraint.

Where Similarity comes in

Each of the theories of categorisation discussed here has at some stage made reference to similarity. As already mentioned, Goodman (1972) is often quoted to make the point that similarity *per se* is too flexible, and that when unconstrained, is meaningless. However, the cognitive system does not treat each new item it encounters as distinct and unrelated from all others (á la Funes). Objects may be judged as being similar to others, and the issue of where and how this relates to concepts and the process of categorisation is one worthy of examination (Hahn & Chater, 1997; Medin, 1989; Medin & Wattenmaker, 1987; Hampton, 1987b; Tversky, 1977).

Similarity and Categorisation – not the same thing

Similar items do not necessarily belong in the same category. Rips (1989) found that even if one item is judged similar to a second, subjects may prefer to categorise it with a third, less similar item. Given a 3-inch round object and asked if it was more similar to a *quarter* (US 25 cents) or a *pizza*, subjects judged it more similar to the *quarter* while still preferring to categorise it as a *pizza*. Keil (1989) performed similar experiments with children, giving them pictures of objects that undergo some transformation. By the age of 8 years, children are certain that a horse painted with black and white stripes is still a *horse* (although more similar to a *zebra*). Rips (1989) also used more involved examples, where subjects were given the story of a bird called a *sorp*, which had the misfortune to live on a radioactive waste dump, causing it to alter over the years. After losing its feathers and growing four extra limbs, this metamorphosis caused the *sorp* to have an insect-like appearance. Subjects were happy to say that the *sorp* was quite similar to an *insect*, but would still only categorise it as a *bird*. Even strong similarity to a different category did not cause these items to move outside their class.

In contrast to the above examples, Keil (1989) also looked at transformations of artifact kinds. When given the example of a *coffee pot* that has been altered to look like a *bird feeder*, the same children were happy to say that the transformed pot was both similar to and had actually become a *bird feeder*. In this case, strong similarity allowed the *coffee pot* to move from one category to another as it was altered, the opposite to judgements with natural kinds. Although often presented as evidence for theory theories of categorisation – i.e. our theories say that artifacts may be transformed while natural kinds may not – these results also serve to illustrate a dissociation between similarity judgements and categorisation. Similarity of items does not necessarily equal categorisation in the same class.

The experiments just mentioned are mainly focussed on context-free categorisation – the canonical forms of *pizza*s, *zebra*s and *bird feeder*s – and moving between different categories. However, Roth & Shoben (1983) found that even with items from the same category, similarity and contextual typicality were not related. For the sentence "the musician tuned the strings of his instrument before playing the classical piece", the most typical exemplar was *violin*, followed by *viola*, *cello*, *bass*, etc., where typicality decreased as similarity to the typical *violin* faded. However, this is not always the case, as seen with the sentence "the square dance musician played his instrument very well". Here, the most typical exemplar is *fiddle*. A *viola*, which is quite a similar instrument to *fiddle* in structure, is in this context judged contextually inappropriate and highly atypical. In contrast, an *accordion* is judged a typical instrument in this context, despite the lack of similarity to a *fiddle*. Other contextual manipulations have been found to produce ordinal reversals of similarities. For example, *grey* is judged to be more similar to *white* than *black* when the context is *hair*, but when the context is *clouds*, the opposite trend is found with *grey* being judged more similar to *black* (Medin and Shoben, 1988). The structure of the category created by context has, like canonical categories, no reliable link with similarity.

Similarity and Categorisation – separate but intertwined

Despite assertions that similarity and categorisation are not the same thing, none of the theories of categorisation are entirely free of similarity considerations. The classical approach, with its necessary and sufficient conditions, is frequently described as rule-based (e.g. Hampton, 1997a), or as similarity-based according to Medin and colleagues (Medin, 1989; Medin & Wattenmaker, 1987; Murphy & Medin, 1985) and thus Komatsu (1992). A novel item is assigned to a category on the basis of its possession of these necessary and sufficient attributes. This process may be viewed as based either on no similarity at all (i.e. a rule), or very constrained similarity (i.e. similar because they share these exact attributes). However, Hahn and Chater (1997) have taken a middle ground and describe classical theories as definitional, with similarity taking a background role. They argue that definitional accounts of categories are not in fact necessary or sufficient, and that with artificial concept learning (Nosofsky, Clark & Shin, 1989; Allen & Brooks, 1991), judgements based on similarity can be seen to intrude even when an explicit rule is present. Unconstrained similarity appears to play a part even when we are following rules.

With family resemblance and exemplar views, the role of similarity is emphasised. A novel item is similar to a family resemblance entity by the weighted attributes they share, or is similar to a group of previously encountered exemplars by a similarity metric that varies with particular exemplar theories. Nosofsky (1986, 1988) even provides a metric of similarity in his Generalised Context Model (an exemplar theory implementation) by measuring it as a function of distance in psychological space. However, similarity again causes problems. As already discussed, unconstrained similarity makes it difficult for the exemplar view to account for category coherence, such as how do we decide what makes a *plum* more similar to a *lemon* than a *lawnmower*? Regarding the family resemblance view, Komatsu (1992) makes the observation that with no a priori constraint on which attributes are considered sharable between instances, there are difficulties setting the boundaries between concepts. Komatsu uses the example of a Bedlington terrier (a dog bred to closely resemble a

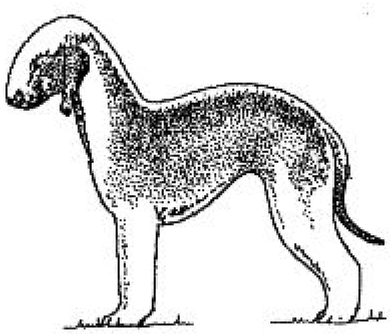

Figure 2.3: a Bedlington Terrier – note resemblance to a lamb

lamb – see Figure 2.3), saying that it appears to share more similarities with lambs than Great Danes (or presumably Chihuahuas). Yet the family resemblance view assumes that its summed weights would lead to it being categorised as *dog* rather than *lamb*. In order to weight the correct attributes to categorise a Bedlington terrier as a *dog*, Komatsu argues that the theory requires the prior partitioning of the world into dogs and lambs. Prior partitioning would therefore have to be the result of a process not similarity-based, and so we have a paradox. Either similarity forms the basis or categorisation, or it becomes redundant.

Final in the list of approaches to categorisation and concepts, theory theory is subject to differing views regarding similarity. It is described as *not* being similarity-based by Medin and colleagues (Medin, 1989; Medin & Wattenmaker, 1987; Murphy & Medin, 1985; also Komatsu, 1992; Hampton, 1997a) as it focuses on the relationships between concepts and our general world knowledge. However, Hahn & Chater (1997) point out that this relation-oriented approach does not undermine the importance of similarity to categorisation, but rather highlights that similarity is not objective. Instead, it may be viewed as being influenced by our theories about how the world works (also Wattenmaker, Nakamura & Medin, 1988). *Beef* and *aubergine* (eggplant) are not usually considered to be similar, yet as part of our theory of what makes an acceptable bulk ingredient in lasagne (where aubergine is a common vegetarian substitute for beef), they do indeed occupy similar roles. Our lasagne theory influences our notion of the similarity between *beef* and *aubergine*.

The ubiquitous nature of similarity in some shape or form is evident across the various theories of categorisation. Rather than abandon similarity in pursuit of various theory theories, many have noted (Medin, 1989; Hahn & Chater, 1997) that there is a greater need for reinterpretation and specification of similarity. Like with theory theory, the main problem is one of constraint. Since similarity and categorisation / concepts are so closely intertwined, constraining similarity will be an important step to the emergence of a well-specified theory of categorisation.

Chapter 3 Co-occurrence Models

Introduction

Data-intensive approaches to semantics are statistical techniques that analyse a set of corpora, and from this derive a summary of the different variety of contexts that different words can be used in. They operate on the principal that if a sufficiently large sample of a language is taken, it can provide useful information about the semantic properties of lexemes in that language and there is a growing body of evidence that supports this. To paraphrase Burgess & Lund (1997), similar words are used in similar contexts, which allows two words to be linked even though they may never appear together.

Co-occurrence Techniques

In co-occurrence analysis, a contextual distribution is calculated for each lexeme encountered in a corpus analysis by counting the frequency with which it co-occurs with every other lexeme (that is, are used together within a particular context, such as a paragraph or moving-window) in the corpora being analysed. The contextual distribution of a lexeme can then be summarised by a vector showing the frequency with which it is associated with the other lexemes in the corpora. One can think of this information as defining a model that contains a network of links between the lexemes in a language, each with varying strengths, thus representing the varying contextual co-occurrence of lexemes in that language. Two such co-occurrence models are the Latent Semantic Analysis (LSA) model (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998), and the Hyperspace Analog to Language (HAL) model (Burgess & Lund, 1997). While the exact parameters of LSA and HAL are different, they both adopt the general approach outlined above to generate co-occurrence vectors.

There is good evidence that co-occurrence analysis extracts information from corpora that can be used to model certain linguistic behaviour. Landauer & Dumais (1997) report that the LSA model can pass a multiple-choice TOEFL synonym test. Lund, Burgess & Atchley (1995) present evidence that co-occurrence data can act as a good predictor of various priming effects. Burgess & Lund (1997) demonstrate that the HAL model can produce clustering in its high-dimensional space of lexemes from differing grammatical categories. The author has chosen LSA as the co-occurrence model to use, because it is conveniently available online (at http://lsa.colorado.edu/), and because of its consideration not only of lexeme-to-lexeme relations, but also relations between a word and its context. By using a large co-occurrence window, the context vector constructed for a word does not (like HAL) take linguistic proximity into account, but rather counts co-occurrence as shared presence in a particular paragraph. When asked to compare two terms, LSA then outputs a similarity score, which is the cosine of the angle between their vectors. This is proportional to Euclidean distance and thus reflects how far apart in high-dimensional context space the points are. The closer the points, the more they have co-occurred with the same contexts and the more similar they will be.

Categorisation and Co-occurrence

Having discussed the importance of typicality, similarity and context to theories of categorisation, it is interesting to note this quote from Rosch (1978):

> "The meaning of words is intimately tied to their use in sentences. …Prototypicality ratings for members of superordinate categories predict the extent to which the member term is substitutable for the superordinate word in sentences. Thus in the sentence 'Twenty or so birds often perch on the telephone wires outside my window and twitter in the morning', the term 'sparrow' may readily be substituted for 'bird', but the result turn ludicrous be substitution of 'turkey', an effect which is not simply a matter of frequency."

Here, she describes typicality in its canonical sense – the more a word is substitutable for its superordinate category name in all its contexts, then the more typical it will be. This notion of substitutability is closely connected with the methodology of co-occurrence techniques. By noting and weighting the surrounding words, the local contexts for a given lexeme is established. Over the entire set of corpora, a typical member name such as *apple* is commonly found to be surrounded by much the same words as its superordinate category *fruit*. In other words, *apple* and *fruit* are found to be rather substitutable for each other. In contrast, an atypical member such as *olive* will not be found surrounded by as many of the same words as *fruit*, and so it is not regarded as substitutable for *fruit* as the word *apple* is. Thus, the closer the distance between the points of a member and its category name, the more typical the member may be.

Regarding similarity in categorisation, it is worth remembering that the original purpose of co-occurrence techniques was to measure similarity of texts for document retrieval. By using the proximity of lexemes' points in high-dimensional space as a measure of their similarity, a co-occurrence model offers a similarity metric that echoes Nosofsky's (1986, 1988) Generalised Context Model (GCM) method of using distance in psychological space. What makes a co-occurrence model different is that each GCM dimension in psychological space represented a feature, for which each exemplar had varying scores. In co-occurrence semantic space, each word is simply represented in terms of other words and does not require explicit hand-coding of feature scores. The distance between two points in LSA semantic space then gives a similarity score for the lexemes. Such a metric of similarity is computationally cheap, making it attractive for use in related research (Ramscar & Yarlett, 2000; Kintsch, 1998). Co-occurrence models such as LSA thus provide a method of measuring (or constraining) similarity that is purely grounded within the language itself.

Finally, the question of context in categorisation also has a foil in co-occurrence models. Each individual lexeme is represented by a meaning vector, occupying a single point in high-dimensional space. In the same way, a sentence (or a paragraph, document etc.) may also be represented by a vector, which will likewise occupy a single point in semantic space. LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a sentence as a kind of average of the meaning of all the words it contains (Landauer, Foltz & Laham, 1998). Thus an LSA similarity score may be given for a word to a sentence as easily as for word to word. As typicality for canonical categories is alignable with the LSA score of an item to its category name, so too is it possible for typicality in context to be alignable with the score of an item to its context sentence.

Revising Theories of Concepts

The functionality of LSA as a co-occurrence model makes it a suitable base upon which to test hypotheses regarding concepts and categorisation. A review of the literature has shown that the belief of concepts as stable encapsulated entities has fallen into disfavour. Certain versions of the theory-based view attempt to deal with this (Barsalou, 1987; Michalski, 1989; Medin & Wattenmaker, 1987; Johnston-Laird, 1983), but these theories and models have so far failed to replicate the success in fitting human data as the exemplar view has (Nosofsky, 1986, 1988).

Recent work by Ramscar & Yarlett (2000; see also Yarlett & Ramscar, 2000) has sought to make explicit this trend towards abandoning encapsulated concepts as the basis for meaningful knowledge representation in psychology. Rather than assuming that human knowledge representations can be formed from fully specified conceptual units, conceptualisation itself is treated as a constructive, contextualised process. Meaningful working representations are built up in context - and in response to task demands - from partial propositional information, retrieved from long-term memory using similarity-based techniques. With regards to categorisation, semantic meaning would not be encapsulated within an object representation, but would instead emerge as the set of relationships between objects in a context-dependent space. LSA, as a co-occurrence technique, operates within this paradigm. Thus, any categorisation tasks modelled would be context-dependent, and based exclusively on the way language is used in the corpora, without explicit hand-coding of category membership or semantic features. The implication is that if LSA – as a statistical tool – can model subject data from experiments hitherto regarded as conceptual in nature, then the question is raised about whether conceptualisation actually plays a role in the processes accessed by these experiments.

Chapter 4 Modelling Categorisation Effects

Introduction

The following section examines a number of hypotheses that test the ability of the co-occurrence model LSA to fit a variety of categorisation behaviours. These include simulations of basic categorisation tasks and modelling the data of typicality experiments. Also, LSA is used to predict subject responses for contextual categorisation, and show that the subjects' typicality ratings in context are different to canonical typicality.

Simulation 1 – Demonstrating Basic Categorisation

The purpose of this initial simulation is to test the ability of LSA to categorise certain semantic categories of words, as demonstrated by the HAL model (Burgess & Lund, 1997). Burgess & Lund used a method of multidimensional scaling to determine if the distances between points were semantically meaningful – i.e. if the points were found to cluster in their category groups. As multidimensional scaling was not available for the LSA model, LSA scores of item against superordinate category name have instead been used as a means of establishing clustering.

Method

A number of words that represented four categories (cities, countries, animal types, body parts) were taken from Burgess & Lund (1997). HAL had found overlap in its context space between items from the city and country lists, and also between the body part and animal lists. To this effect, the simulation was split into two parts, to test for independence of the category pairs:

- city and country
- body part and animal type

Each category's data set was revised (to remove items such as *Africa* from the country list), and increased to offer a greater number of items (20-22) per category. For each pair of categories, all items were compared in turn to both superordinate category names (*city*, *country*), (*animal*, *body*) and the similarity scores noted. All scores were calculated in LSA using the General Reading up to 1st Year College semantic space, with term-to-term comparison and maximum factors. All materials and scores are available in Appendix A.

Results

The categorisation scores for each of the category pairs are shown in the graphs below, where the axes are LSA scores for each superordinate category name in the pair. Figure 4.1 shows the distribution for cities (denoted by filled triangles) and countries (denoted by open circles). The x-axis represents the LSA score that every

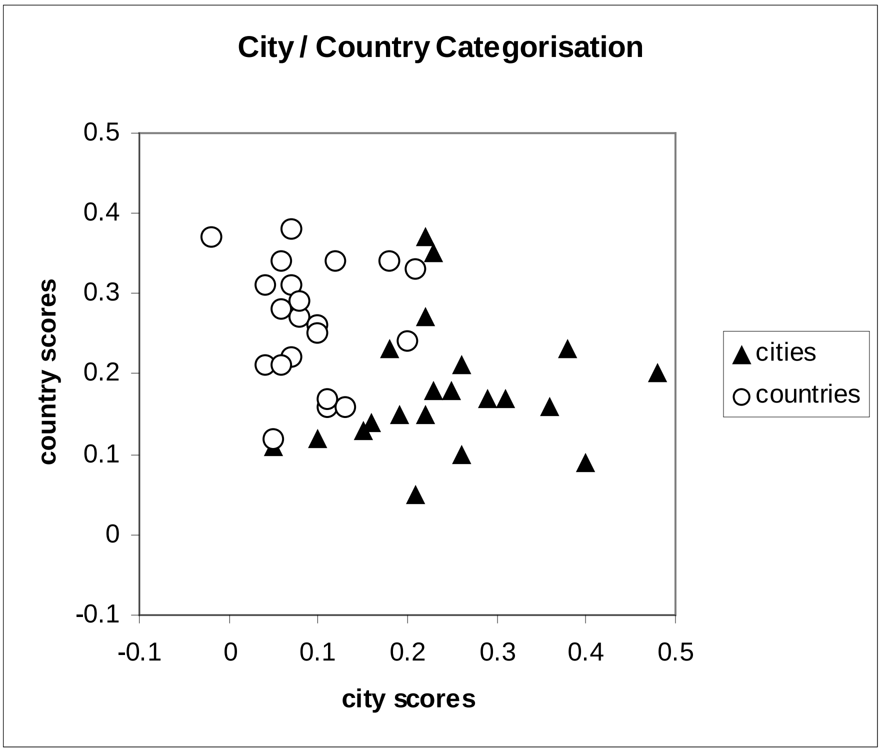

Figure 4.1: graph of city / country categorisation

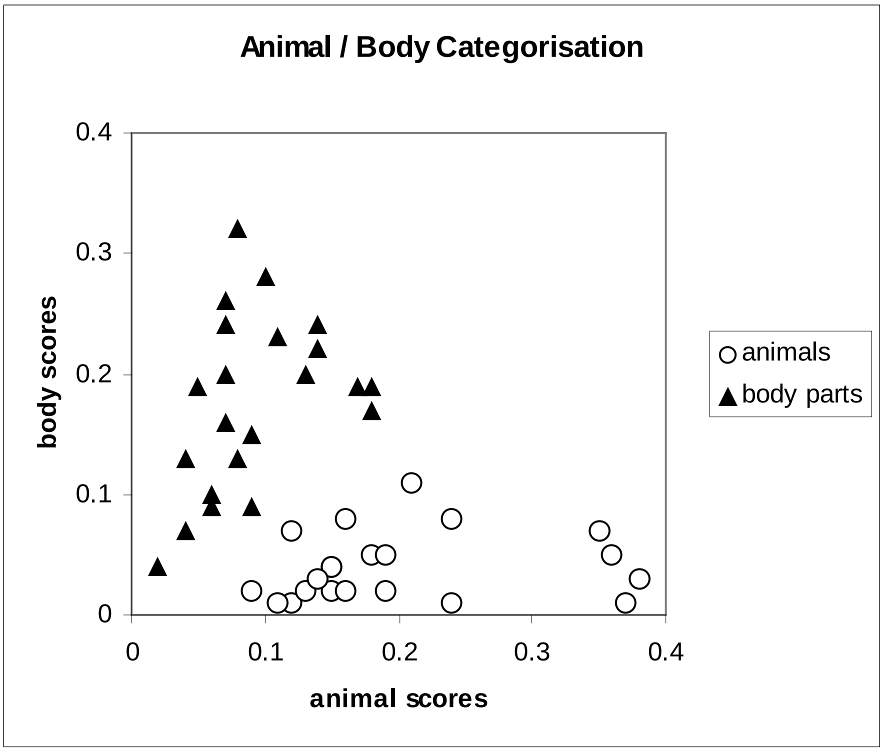

Figure 4.2: graph of animal / body categorisation

item in the pair set received against *city*, and the y-axis represents the scores against *country*. Figure 4.2 then shows the clustering for animal types (denoted by open circles) and body parts (denoted by filled triangles). The x-axis represents the LSA score that every item in the pair set received against *animal*, and the y-axis represents the scores against *body*.

Items in the set of cities were significantly differentiated from countries, when scored against *city* (t=6.13, df=40, p<0.001), and *country* (t=3.61, df=40, p<0.001). Likewise, the differentiation of animal types from the set of body parts was significant, for scoring against both *animal* (t=4.55, df=38, p<0.001), and *body* (t=8.21, df=38, p<0.001). Independent 2-tailed t-tests, assuming equal variances, were used in each case. Some items were found to be more closely surrounded by members of the opposite set – e.g. Sydney and Melbourne (city set – visible as the cities with the highest LSA score against *country*), and Mexico (country set – visible sandwiched between two cities near the middle of Figure 4.1). The (animal, body) category pairs were more cleanly divided into clusters, although two items from the body part set scored quite low against both category names and can be found at the bottom-left corner of Figure 4.2 – tooth and eyelid at (0.02, 0.04) and (0.04, 0.07) respectively.

Discussion

Given that words with similar meanings tend to be close to each other in LSA's high-dimensional context space, we can agree with Burgess & Lund (1997) and Laham

(1997) that co-occurrence vectors from the high-dimensional context space carry information that mimics semantic knowledge. This can then be used to carry out simple categorisation tasks that may divide members from non-members of a category. The nature of shared category membership can be seen with the item Mexico in the (city, country) graph, where it received a score (0.20, 0.24). Since Mexico is both the name of a country and its capital city, its original inclusion in the set of countries (as opposed to cities) was arbitrary. Its score shows that it is similar to both *city* and *country*, though lying closer to *country*, and thus may be considered a member of both categories. Category members cluster around their superordinate category name (proximity being alignable with substitutability in the corpora) and likewise, keep their distance from other category names.

It is interesting to note that the two items from the cities list that were found mainly surrounded by items from the opposite category (countries) were Melbourne and Sydney, the only two Australian cities in the set. This is because the context vectors created for these two cities were more similar to *country* than *city*. As conjecture, this may be because the corpora used in LSA are American texts, where discussion of Australian cities may not be as contextually diverse as if they were American cities, or as if the corpora themselves were Australian. To further support this, the five items from the city list that lie furthest from the cluster of countries are mostly American cities. Visible on the graph in Figure 4.1 as the cities that received the highest LSA scores against *city*, these are Chicago, Seattle, Miami and Atlanta, the non-American exception being Tokyo. This is an indication of the influence that corpora choice has on the context space. LSA may be seen as a co-occurrence model of American English, and having used American corpora (made up of texts, novels, newspaper articles, etc.), has a context space with an American “perspective”. It follows that the context vectors for little-discussed words (such as Australian cities) will not be as informative as words that are frequently and widely used (such as American cities).

From this simulation, LSA has been shown to categorise items belonging not just to concrete categories (animals, body parts) but also to more abstract category types (cities, countries). There is little difference in performance as there is no distinction between these types of noun – or indeed between any other lexemes – in LSA. Also, considering the number of contexts that cities and countries share – as an indication, *city* against *country* receives an LSA similarity score of 0.27 (parameters as before) – the fact that these two categories were so clearly separable serves as an extra highlight of the subtlety of high-dimensional context space. Although dependent on the diversity and size of the corpora, co-occurrence techniques can extract enough semantic information to perform simple categorisation tasks, without presupposing primitive or defining semantic features or requiring an experimenter to commit to a particular type or set of features.

Simulation 2 – Demonstrating Typicality in Categorisation

Having shown that LSA succeeds in categorisation tasks for both concrete and abstract categories, the next stage is to test the model’s ability to deal with canonical typicality effects. The purpose of this simulation is to use data from typicality studies (Rosch, 1973; Armstrong, Gleitman & Gleitman, 1983; Malt & Smith, 1984) to see if a correlation exists between subject typicality scores and LSA similarity scores for members against their superordinate category name. Categories that were common to two or more studies also had the typicality scores of their shared items compared, to see if the LSA scores fell within the range of inter-group differences.

Method

Each set of typicality data was divided up according to its original set:

- Set A from Rosch (1973)
- Set B from Armstrong, Gleitman & Gleitman (1983)
- Set C from Malt & Smith (1984)

Within these three data sets, 18 sets of typicality ratings existed, across 12 separate categories. Set A and Set B had 4 categories in common, where all Set A's items were present in the larger Set B. However, while Set C shared 2 and 1 categories with sets A and B respectively, there were not enough common items for a valid rank correlation of typicality scores.

For each category in each data set, all items were compared to the superordinate category name and the similarity scores noted. All scores were calculated in LSA as for Simulation 1, using the General Reading up to 1st Year College semantic space, with term-to-term comparison and maximum factors.

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies, where a score of 1 represents the most typical rating. Malt & Smith used the 7-point scale in reverse order (where 7 represented most typical) so these scores were inverted. LSA score scaling was done by aligning the highest of the LSA scores for each category with the most typical rank on the 7-point scale; i.e. the highest LSA score for a category would be matched to 1, and the other scores falling proportionately towards 7. The exact formula is given in Appendix E. Full tables of materials and scores are available in Appendix B.

Results

Spearman's rank correlation (*rho*) was used to compare scaled LSA and subject scores. The global rank correlation between the subject ratings and LSA scores across Sets A, B and C (193 items) was *rho* = 0.515 (p<0.001). Table 4.1 shows these correlation coefficients with the level of significance (*p* for 2-tailed test) for each of the 18 rating sets. Many of the categories that failed to produce greatly significant correlations benefited from the removal of one member, due to it having an extremely high or low LSA score. The new rank correlation coefficient and level of significance for these adjusted sets can also be seen in Table 4.1; where there is no adjusted score given, the category did not benefit significantly from the removal of an item. To compare with LSA's performance in fitting subject data, Set A scores were correlated with those of the relevant items in Set B for the 4 shared categories. The new *rho* and levels of significance for these categories are also shown in Table 4.1, within the Set A band. However, any level of *p* below 0.10 was considered insignificant and omitted from the table.

Table 4.1: Rank correlation coefficients *rho* (with levels of significance *p*) between LSA and subject scores and between Set A / Set B where applicable

	Category	Rank correlation *rho* (level of significance *p*)		
		initial	adjusted	Set A / Set B
Set A (Rosch, 1973)	sport	1.000 (0.01)		1.000 (0.01)
	fruit	0.886 (0.05)		0.943 (0.05)
	vehicle	0.829 (0.10)	1.000 (0.05)	0.886 (0.10)
	crime	0.814 (0.10)	0.975 (0.10)	
	bird	0.714 (0.10)	0.900 (0.10)	
	science	0.414 (-)	0.675 (0.10)	
	vegetable	0.371 (-)		0.886 (0.10)
Set B (Armstrong, Gleitman & Gleitman, 1983)	sport	0.811 (0.01)		
	vehicle	0.788 (0.01)		
	vegetable	0.580 (0.10)	0.745 (0.05)	
	fruit	0.539 (0.10)	0.748 (0.05)	
	female	0.346 (-)	0.558 (0.10)	
Set C (Malt & Smith, 1984)	trees	0.705 (0.01)		
	clothing	0.521 (0.05)	0.676 (0.05)	
	furniture	0.466 (0.05)	0.609 (0.01)	
	bird	0.375 (-)	0.640 (0.05)	
	fruit	0.157 (-)		
	flowers	-0.499 (-)		

Values shown as (-) represent insignificant correlation

It must be noted that the same rank correlation coefficient results in differing levels of significance within the table. This is due to different sizes in categories' data sets (from 5 to 20), where the same score could be significant for one size set and not another; e.g. perfect rank correlation of 1.000 is significant to $p < 0.01$ with N=10, but only to $p < 0.05$ when N=5. Likewise, rank correlation of 0.609 is significant to $p < 0.01$ when N=20, but would not be for a smaller N. This high sensitivity to the degrees of freedom from small-sized data sets is why one item was capable of skewing the rank correlation (as shown above). With small data sets such as these, the power of the tests being used is restricted and they are overly sensitive to individual data points. Thus, it seems reasonable to consider as marginally significant those results where $p < 0.10$, given the constraints of the data.

.

Discussion

In this simulation, LSA similarity scores correlate significantly with subject typicality ratings, and thus support Rosch's statement that typicality may be considered a measure of canonical substitutability. Having demonstrated basic categorisation tasks in Simulation 1, we now see that LSA's semantic space can also model gradient of typicality within a category. Significant global correlation existed between LSA-to-subject typicality ratings at *rho* = 0.515 (*p*<0.001, *N*=193). Items that subjects judged typical correlated with those that LSA scored highly in similarity with the category name. The same correlation is true of items that subjects judged to be highly atypical members of their category – these received low similarity scores in LSA. The more closely the ranking of LSA scores mirrored that of the subjects', the higher the correlation, and the closer the level of significance (*p*) dropped to zero. Only one score of perfect correlation 1.000 was found without adjustment, for the category *sport* in Set A (see Figure 4.3).

Of the 12 separate categories across the data sets, the only typicality gradient that LSA failed to model to any significance is that of *flowers* (Set C), which actually scores negative correlation of –0.499. This could be the result of the nature of the texts that make up the corpora used in this semantic space, that contain little contextual variation for the category and flower names used. In particular, LSA scored unusual items such as *poinsettia* as equally or more similar to *flowers* than *daisy* or *rose*. If one notes the vector length for *poinsettia* in LSA, it is given as 0.03, as opposed to 0.74 for *rose* and 0.23 for *daisy*. This would result if limited texts were available to build up the relevant context vectors for *poinsettia*, which may only have one or two occurrences in the entire corpora. Even “intuitively” one can imagine daisies and roses occurring in many contexts where one would not expect to find poinsettia. It is possible that larger, more balanced corpora would amend this quirk, or contextual blind spot, but for a low-frequency word like *poinsettia* this cannot be guaranteed.

Regarding the other 11 separate categories, there were cases where LSA modelled a category’s typicality gradient successfully in one data set but not in another. An example of this is the category *fruit*, which was modelled with rank correlation of 0.886 ($p < 0.05$) in Set A and 0.748 ($p < 0.05$) in Set B (adjusted), but failed to correlate significantly at all in Set C. Another case is that of *vegetable*, which was modelled for Set B with correlation of 0.58 ($p < 0.10$), but failed to correlate significantly for Set A. The removal of the item *carrot* led the Set B correlation with LSA to rise to 0.745 ($p < 0.05$). Similarly, the removal of the item *penguin* from Set C’s category *bird* brought the insignificant correlation coefficient of 0.375 to 0.640 ($p < 0.05$). Again, the unusually high similarity score of *penguin* to *bird* in LSA semantic space could be attributed to the particular choice of corpora used. A more balanced corpus should contain many contexts for *bird* that would cause its co-occurrence vector to differ greatly from that of penguin. In its current state, *penguin* appears to be another quirk in LSA semantic space. This was the reason that many sets of ratings (10 out of 18) benefited significantly in correlation with the removal of one item. The LSA scores for these quirky items caused the correlations suffer.

The small number of items per category – six – in Set A (Rosch, 1973) made it difficult for the correlations with LSA to gain significance. An example of this may

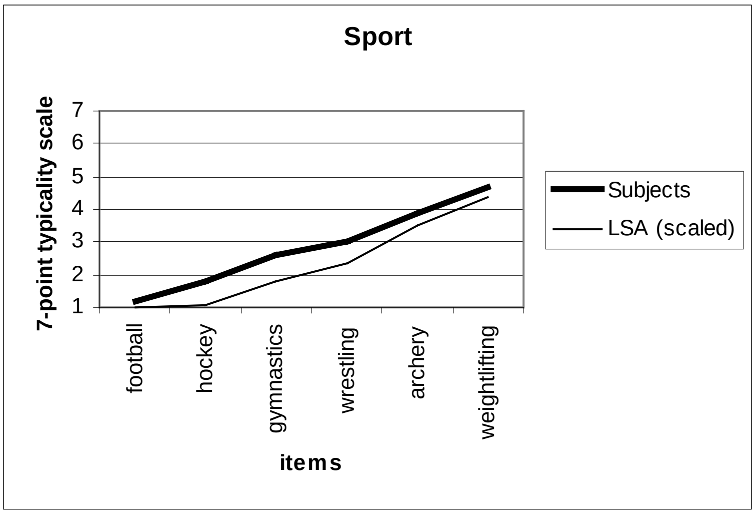

Figure 4.3: graph of Rosch (1973) and LSA scores for category *sport*

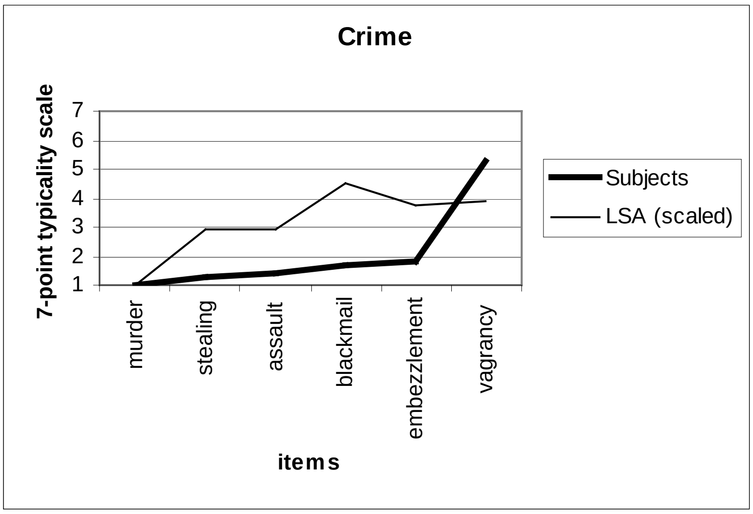

Figure 4.4: graph of Rosch (1973) and LSA scores for category *crime*

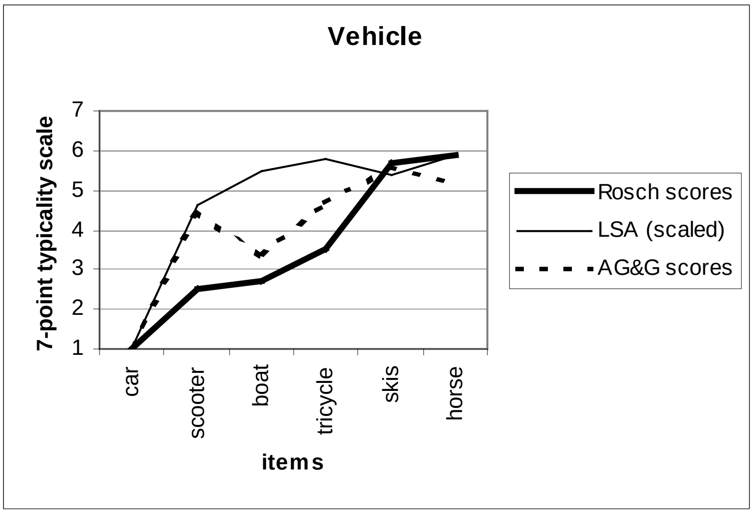

Figure 4.5: graph of Rosch (1973), Armstrong, Gleitman & Gleitman (1983) and LSA scores for category *vehicle*

be seen in Figure 4.4, a graph for the category *crime*. Initial rank correlation was 0.814 (significant to p < 0.10), but removal of the item *blackmail* made this rise to 0.975, which is still only significant to p < 0.10. Larger category data sets are to be found in Sets B and C, where although the rank correlation coefficients may be lower, they are more significant. If Rosch had used even slightly larger category data sets, the correlation coefficients would not be quite so sensitive to the degrees of freedom.

Only one of the 5 category types in Set B came from what Armstrong, Gleitman & Gleitman (1983) term as well-defined categories – the category *female*. Other definable categories from their experiments (such as *odd number*, *even number*) were unsuitable for use in LSA, and are instead analysed in Simulation 3. With the adjustment of the category *female* to remove the item *chairwoman*, rank correlation rose to 0.558, making a significance level of p < 0.10. This was one of the worst performances on LSA's part to still achieve some level of significance (however tenuous). It is unclear why Armstrong, Gleitman and Gleitman chose to regard *female* as a well-defined category. If the "rule" by which they claim to define *female* is by gender, then it is also arguable that *fruit* or *vegetable* (or any other taxonomic category) may be defined by its genus. It seems reasonable to regard typicality in *female* as one would any other category examined in this simulation – a measure of contextual substitutability. In this case, the contextual substitutability shown by LSA similarity scores failed to convincingly model the typicality scores for *female*, only reaching 0.10 significance when the category was adjusted. A possible reason for this is that typicality ratings for a category such as *female* are subject to social conditioning in a way other categories such as *fruit* or *sport* are not. For example, the item that LSA scored highest against *female* was *housewife*, which was next followed by *chairwoman*. Although this simply reflects the general contextual substitutability of the words across all of LSA's corpora, it also reflects a ranking that may not be found within a social group. It would be inconsistent for a group of subjects to rate *housewife* as the most typical *female* (a stereotyped sexist attitude), while rating *chairwoman* (a stereotyped politically correct attitude) closely behind. Thus LSA may have failed to convincingly model this category's typicality gradient because it reflects a variety of social attitudes across its corpora, and not just those of 1980's Philadelphia undergraduates

One of the most interesting findings is that in 3 out of 4 cases of shared categories between Set A and Set B, LSA provided as good a fit to Set A typicality ratings as Set B did. When the item *skis* was removed from Set A's *vehicle* category, LSA's correlation bettered that of Set B (see Figure 4.5 for the initial graphs). The exception to this performance was the category *vegetable*, as already discussed. This serves to make an important point and put the data in Table 4.1 into perspective: it suggests that the difference between subject groups in Rosch's (1973) and Armstrong, Gleitman & Gleitman (1983) experiments is comparable to the difference between LSA and human subjects. In other words, a co-occurrence model like LSA is as successful at matching the typicality gradients of a subject group as another subject group would be.

Simulation 3 – Demonstrating Typicality in Well-Defined Categories

As mentioned in Simulation 2, of the categories described as "well-defined" by Armstrong, Gleitman & Gleitman (1983) are not all suitable for modelling in LSA. Only *female* was used, and has already been discussed. The remaining three well-defined categories are *odd number, even number* (unsuitable for modelling in LSA because of its use of digits) and *plane geometry figure* (unsuitable because of its perceptual rather than linguistic nature). However, these categories were still ascribed a typicality gradient by subjects, and the purpose of this simulation is to offer a possible explanation for typicality in the categories *odd number* and *even number*, by demonstrating its correlation with simple frequency rather than being a function of categorisation.

Armstrong, Gleitman & Gleitman (1983), despite using frequency norms from Battig and Montague (1969) for the categories they considered "prototypical", applied no such constraint on the items they selected for the categories *odd number* and *even number*:

> "Since there are no previously collected norms for the well-defined categories
> we used here, two sets of six exemplars were generated for each category on
> the basis of an intuitive ranking made by the experimenters."

Although Armstrong, Gleitman and Gleitman later state that frequency counts for some numbers are available in Kucera and Francis (1967), they only examine the issue of frequency with regards to reaction times in their sentence verification experiment. Word frequency is not examined by Armstrong, Gleitman and Gleitman for the items and typicality scores taken from their rating experiments for use in this simulation.

Method

Using the British National Corpus (BNC), a frequency count was established for each of the numbers used in the categories *odd number* and *even number* from the Armstrong, Gleitman and Gleitman (1983) experiments. The full BNC (over 100 million words) was used. Only the numeric form of the numbers was counted rather than the alphabetic (e.g. counting only occurrences of “3”, and not “three”).

Results

The frequency count for each number in both categories is shown in Table 4.2 beside its typicality scores from Armstrong, Gleitman and Gleitman (1983). Spearman’s rank correlation for BNC frequencies to subject scores was –0.891 (p < 0.01) for *odd number*, -0.920 (p < 0.01) for *even number*, and –0.939 (p < 0.01) for both categories combined. Correlation is negative because typicality is rated on a descending scale (most typical rating is 1) while frequency counts ascend.

Table 4.2: BNC frequencies and subject typicality scores for “well-defined” categories *odd number* and *even number*

Odd Number	Subject score	BNC frequency	Even Number	Subject score	BNC frequency
3	1.6	25040	2	1.0	34394
7	1.7	10676	4	1.1	20071
11	1.7	9238	8	1.5	9713
13	1.8	7779	10	1.7	18570
9	1.9	7954	6	1.7	13425
23	2.4	4771	18	2.6	8077
57	3.0	756	42	2.6	1415
501	3.5	57	1000	2.8	963
91	3.7	442	34	3.3	1496
447	3.7	31	106	3.9	259
			806	3.9	38

Discussion

The typicality gradient in a category such as *odd number* and *even number* correlates with the frequency of occurrence of the numbers in a representative corpus. This suggests that what is happening in a task such as Armstrong, Gleitman and Gleitman’s (1983) is not the result of categorisation and/or conceptual processing. Rather, in response to an artificial laboratory environment, subjects rated numbers with typicality according to their general frequency of occurrence.

Having already questioned the description of *female* as a well-defined category, this alignment of typicality for *odd number* and *even number* with mere frequency puts any discussion of well-defined categories on an unsteady footing. LSA may be used to model the typicality of categories like *female* in terms of co-occurrence in a high-dimensional context space, and simple frequency counts may be used to model the typicality of categories like *odd number* and *even number*. The effect of this is that Armstrong, Gleitman and Gleitman’s arguments for a distinction between well-defined and fuzzy concepts has lost its impetus, and that in fact there is no such thing as a well-defined concept. As already discussed, the category *female* is not well-defined, but as fuzzy as *fruit* or *sport*. Rather than considering *odd number* and *even number* in terms of context-space distance from a category name, an even simpler mechanism can account for subjects' tendency to assign them typicality gradients. The typicality of items in *odd number* and *even number* comes about from the plain effects of frequency, and thus has little to do with conceptual thought.

Simulation 4 – Demonstrating Basic Context Effects

The first two simulations have shown that a co-occurrence model such as LSA can be used to demonstrate basic categorisation tasks and typicality judgements in canonical (context-free) categories. However, categorisation is also subject to linguistic context, whose capacity of to skew typicality has been demonstrated by Roth & Shoben (1983). The purpose of this simulation is to test the use of LSA's similarity score as a metric of "relatedness", to model how context and contextual relatedness affects subject reaction time.

Method

Roth and Shoben had asked subjects to decide whether an item was a possible referent of a category term in a context sentence: e.g. "Stacy volunteered to milk the animal whenever she visited the farm" with the item *goat* (see Table 4.3). One of 4 possible items (category members) was given per sentence. Two of the possible items were true in that context (*cow*, *goat*), and two were false – i.e. "impossible" (*bull*, *bear*). Of the true items, one was chosen to be closer to the context than the other (*cow* = related, *goat* = unrelated). The same was done with the false items, (*bull* = related, *bear* = unrelated). In this sense, relatedness is alignable with LSA similarity score. Roth and Shoben found that the reaction times (RT) of subjects depended on the relatedness of the item to the context. i.e. The more related an item was to the context sentence, the easier it was for subjects to confirm the sentence if it was true. However, the more related an item was to the context sentence, the more difficult it was for subject to reject the sentence if it was false. In other words, subjects were faster to agree to *cow* than *goat* in the context of milking an animal on a farm, but slower to reject *bull* in this context than *bear*.

The context sentence was compared in LSA to each referent item. Owing to the earlier observed quirk in LSA semantic space that led to *penguin* scoring extremely highly against *bird*, the step was taken to omit the category name from the sentences for each comparison. The similarity scores were noted, then grouped by true and false items. All scores were calculated in LSA using the General Reading up to 1st Year College semantic space, with document-to-term comparison and maximum factors.

Table 4.3: Referent items for two example context sentences

		Sentence 1: Stacy volunteered to milk the animal whenever she visited the farm	Sentence 2: The hunter shot at the bird flying high overhead
True	Related	cow	duck
	Unrelated	goat	crow
False	Related	bull	chicken
	Unrelated	bear	penguin

This is an illustrative simulation only; since Roth & Shoben did not make their data set available, only two full examples could be used.

Results

Table 4.4: LSA scores for referent items against
context sentences 1 and 2 (from Table 4.3), with reaction times (RT)

		Sentence 1 LSA	Sentence 2 LSA	Mean LSA	RT (msec)
True	Related	0.59	0.30	0.45	1144
	Unrelated	0.32	0.30	0.31	1747
False	Related	0.14	0.24	0.19	1496
	Unrelated	0.06	0.25	0.16	1293

Table 4.4 shows the mean RT and LSA scores for both of the context sentences and their true / false referent items. Pearson's *r* correlation between mean LSA score and mean subject RT for true items was -1.00 ($p < 0.01$), i.e. faster RT corresponded with higher LSA similarity scores (higher relatedness). For false items, the correlation was 1.00 ($p < 0.01$), i.e. faster RT corresponded with lower similarity scores (lower relatedness).

Discussion

LSA's metric of similarity between a sentence and a possible referent correlates with subjects' RT to sentence verification tasks. Although all possible referents per sentence were members of the same category, the false items made the sentence "impossible" (such as milking a *bull* / *bear*, or a *chicken* / *penguin* flying overhead). Subjects took longer to reject a false sentence if the item was related to the context (such as a *bull* with a farm). Using LSA's similarity scores as a measure of relatedness in the context is quite straightforward. The co-occurrence vector created in LSA for each sentence occupies a point in high-dimensional context space. The distance between this point and that of a referent item is a measure of similarity between the two. Thus *cow* was closer to the context sentence than *goat*, so the degree of relatedness was higher. LSA similarity scores can capture this degree of relatedness, and thus can reflect the RT required to process the decision. For false sentences, referent items with higher LSA scores will have longer RT. For true sentences, referent items with higher LSA scores will have faster RT.

Detecting this relatedness is a demonstration of LSA's contextual sensitivity. Having already modelled categorisation tasks and canonical typicality, this simulation deals with a more complex issue. Rather than just considering canonical categorisation and typicality as a function of contextual substitution, the presence of a context can also be considered in terms of similarity. Roth & Shoben (1983) saw the context sentence as something that constrains the categorisation process, and alters the structure of the category in response. Termed the restructuring hypothesis, they describe it using a spatial analogy:

> "Context can be thought to shift the focus point in the space to some new point that represents the attributes suggested by the context. This point would not necessarily correspond to a particular exemplar. …Goodness-Of-Example would be a function of distance"

Although Roth and Shoben make reference here to attribute space (bringing to mind Tversky's (1977) Contrast Model) their description can still be highly compatible with co-occurrence models. We must think instead of the focus point as representing a point in high-dimensional context space – a context vector. For canonical categorisation (before context is introduced), the focus point can be thought of as the category name. Once context is introduced, we are looking at contextual categorisation and the category is restructured so that the new focus point is the context vector created for the whole sentence. In both cases, Goodness-Of-Example – or typicality as it is more often called – is the distance from this focus point and is expressed as a similarity score in LSA.

In this simulation, the ability of LSA to capture some of the more subtle phenomena of categorisation tasks has been illustrated. Roth and Shoben concluded from their experiment that once context is introduced, typicality, as determined in isolation, no longer plays an important role. The author agrees with this, showing that LSA can be used to model not only canonical typicality, but also the effects of context where category structure can be altered significantly.

Experiment 1 – Typicality in Context

Having demonstrated both canonical typicality in Simulation 2 and the use of co-occurrence vectors to represent context in Simulation 4, the purpose of this experiment was to test if LSA could be used to predict subject responses for typicality in context. The hypothesis was that LSA could predict human judgements of exemplar appropriateness (typicality) for given context sentences. LSA similarity scores were used for each context sentence to form significantly different clusters of appropriate (high scores) and inappropriate (low scores) items. The anticipation was that the subject ratings of typicality in context for these items would fall into the same clusters, and that these clusters would also be significantly different.

LSA Method

Materials consisted of 7 context sets, each of which consisted of a context sentence and 10 possible members of the category. 3 of the context sentences were taken from Roth & Shoben (1983), the other 4 by the experimenter. Category members were chosen in two ways, to form the appropriate and inappropriate clusters for the context.

Firstly, appropriate items were found by taking 4-5 category members that appeared in the LSA list of 1500 near neighbours of the context sentence. This list corresponds to the 1500 points in LSA's high-dimensional space that are closest to the context sentence, and would receive the highest similarity scores. The sentence was processed as a pseudodoc using maximum factors in the General Reading up to 1st Year College semantic space, from which all words in the corpus with a frequency of less than or equal to 5 had been removed.

Secondly, inappropriate items were found by compiling a large list of category members and taking 5-6 of those with the lowest (preferably negative) LSA similarity score against the context sentence. The scores were calculated in LSA by comparing the context sentence to each item in the list, using the General Reading up to 1st Year College semantic space, with document-to-term comparison and maximum factors.

An example of one such context set with clusters and LSA scores is given in Table 4.5.

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies. This was done by aligning the extremes of the LSA scores for each category with the opposite extremes of the 7-point scale; i.e. the highest LSA score for a category would be matched to 1, the lowest score to 7, and the intermediate scores falling proportionately in between. The exact formula is given in Appendix E. All materials and scaled LSA scores are available in Appendix C.

Table 4.5: Sample context set with
appropriate / inappropriate clusters and LSA scores

Context Sentence	Appropriate items	Highest LSA scores	Inappropriate items	Lowest LSA scores
Fran pleaded with her father to let her ride the [animal]*	mare	0.38	cow	0.02
	stallion	0.31	elephant	0.01
	pony	0.27	dog	-0.01
	horse	0.26	bear	-0.02
	mule	0.17	tiger	-0.03

* [animal] was not present in the sentence and marks the position where an item was placed

One point to note is that LSA is sensitive to the personal names used in the context sentences. These were originally picked arbitrarily (except where taken directly from Roth & Shoben, 1983), but closer examination proved that changing the names could change the LSA similarity scores by 0.01-0.05. However, the ranking of the items rarely changed significantly, and the cluster assignment never did so, and the experiment was run with the original choice of names.

Subject Method

Subjects

The subjects were 19 native speakers of English and were all volunteers who participated remotely as part of a web-based experiment.

Materials

Materials were as for the LSA method (7 context sets with 10 possible category members), split into two sections. Each section consisted of 7 context sets, now with 5 items, selected so that there were at least 2 of both appropriate and inappropriate items in the set. Where there was a pair of context sets focussed on the same category in different contexts, any item was common to both sets was only used once per section. The sections were alternated for each subject. All 35 items within each section were randomly presented for each subject. All materials and mean subject ratings are available in Appendix C.

Procedure

Subjects read instructions that explained typicality and the 7-point scale as per Rosch (1973) and Armstrong, Gleitman & Gleitman (1983) (See Appendix F). They were then given this example of a context sentence (not used in experiment) “The girl played the GUITAR while the others sang around the campfire”, and told to consider the appropriateness of the capitalised word in the context given.

Subjects were asked not to spend more than 10 seconds deciding on what score to give, and were warned that it would not be possible to go back and change an answer (owing to the software used). Any problems or comments could be e-mailed to the experimenter. Steps were taken to ensure that each subject only provided one set of typicality ratings.

Results

Subjects agreed with LSA's predictions of typicality for 10/10 items in 3 context sets, for 9/10 items in 3 further context sets, and for 5/10 in the remaining 1 context set. Significant difference in clusters, not rank correlation, is the important factor here, because even subject data with low correlation to the LSA score may fall into the two specified clusters (and thus prove the prediction hypothesis true).

For all 7 context sets, Mann-Whitney (Wilcoxon Summed Ranks, 2-tailed) test showed the LSA scores fell into two significantly different clusters. The results varied when testing with subject scores for difference between the predicted clusters, from three context sets having significant differences at $p < 0.01$ (those at 10/10 agreement), to one set failing to achieve any significant difference at $p = 0.69$ (5/10 agreement). Data for clustering in both LSA and subject scores are given in Table 4.6. Three of the context sets that only produced clusters that were significantly different to $p < 0.10$ were those where subjects agreed with LSA-predicted clusters for 9/10 items. With the removal of this lone contentious item, each of these three adjusted subject sets achieved significance of $p < 0.05$ (actually $p = 0.016$), and these results may also be seen in Table 4.6. Reasons for considering significance of $p < 0.10$ are the same as those for Simulation 2, stemming from small data set size.

Table 4.6: Context sentences for LSA, subject scores, adjusted sets giving Wilcoxon's *W* and significance of difference between clusters

Context Sentence	LSA scores		Subject scores		Adjusted subject sets	
	W	Significant	*W*	Significant	*W*	Significant
Stacy volunteered to milk the [animal] whenever she visited the farm *	10	$p < 0.01$	10	$p < 0.01$		
Fran pleaded with her father to let her ride the [animal] *	15	$p < 0.01$	15	$p < 0.01$		
The [bird] swooped down on the helpless mouse and carried it off	10	$p < 0.01$	10	$p < 0.01$		
Jane liked to listen to the [bird] singing in the garden	15	$p < 0.01$	18	$p < 0.10$	10	$p < 0.05$
Jimmy loved everything sweet and liked to eat a [fruit] with his lunch every day	15	$p < 0.01$	18	$p < 0.10$	10	$p < 0.05$
Sophie was a natural athlete and she enjoyed spending every day at [sport] training	15	$p < 0.01$	19.5	$p < 0.10$	10.5	$p < 0.05$
During the mid morning break the two secretaries gossiped as they drank the [beverage] *	15	$p < 0.01$	25	$p < 0.70$**		

* Sentences taken from Roth & Shoben (1983)
** Not significant but included for completeness

Discussion

The results support the basic hypothesis that, in the majority of cases, LSA can predict whether members of a category will be appropriate or inappropriate in a given context. In other words, LSA can predict human judgements of typicality in context as well as in canonical categories (as demonstrated in Simulation 2). For example, LSA predicted in the context set for *animal* ("Fran pleaded with her father…") that the item *elephant* would be placed in the inappropriate cluster, even though it is entirely possible to ride on an elephant. A problem with subject disagreement had been anticipated because of this. However, the prediction was consistent with subjects' judgements, where *elephant* received a typicality score of 4.1 and resided in the inappropriate cluster. Is this respect, LSA predictions were sometimes unexpectedly accurate.

In 3 of the 7 context sets, subject typicality scores agreed with LSA predicted clusters for 10/10 items and separated the clusters to a difference significance of $p < 0.01$. These sets involved natural kinds as the category for which typicality was taken (*animal*, *bird*). In a further 3 context sets, subjects agreed with LSA's clustering for 9/10 items and separated the clusters to a significant difference of $p < 0.05$ when these 9 items were considered. For these sets, two categories were of natural kinds (*bird*, *fruit*) and one was an abstract artifact kind (*sport*). Finally, the context set for which only 5/10 items were agreed to be in the predicted clusters was also for an artifact kind (*beverage*). This suggests that LSA may perform better in predicting the contextual typicality of natural kinds than artifact kinds.

The *beverage* context set ("During the mid morning break…") was the only one of the seven that failed to produce any significant clusters for subject scores (W=25, $n1$=$n2$=5, p<0.70), and was an artifact kind. Of the 10 items that were predicted to cluster into 5 each per appropriate and inappropriate clusters, subjects agreed with 1 of the predicted appropriates (*coffee*) and 4 of the inappropriates (*water*, *cola*, *cocoa*, *saki*). The remaining 5 items were each allotted to the opposite cluster to that which LSA had predicted, with subjects allowing only 2 appropriate items (*coffee, tea*) and scoring all 7 others as inappropriate (see Figure 4.6). As mentioned above, LSA appeared to perform better with natural kind categories than artifact kinds, perhaps as

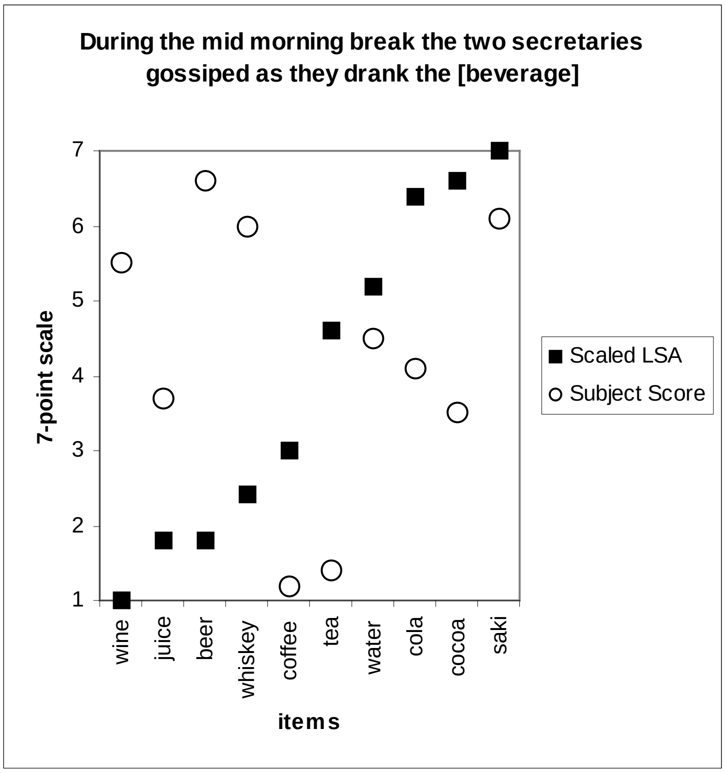

Figure 4.6: graph of scaled LSA scores and subject ratings for *beverage* in context (agreement on 5/10 items – only context set that failed to cluster significantly)

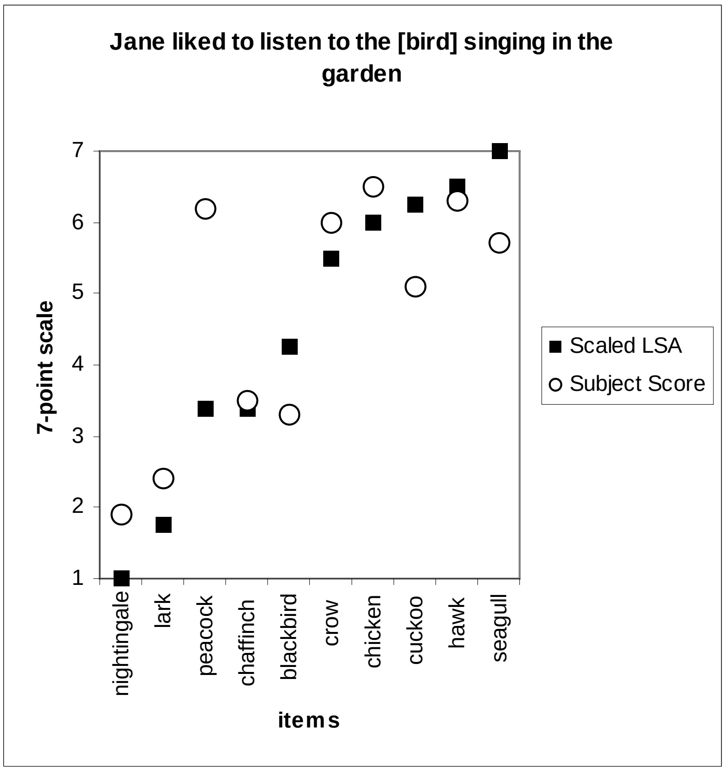

Figure 4.7: graph of scaled LSA scores and subject ratings for *bird* in context (agreement on 9/10 items – clusters significantly at *p*<0.05 without *peacock*)

a result of the vectors for artifact kinds containing a greater degree of contextual variation and thus scoring more unpredictably against the context sentence. This is compatible with psychological data showing that artifact kinds are processed differently because they may be found in a variety of functional and relational roles, and/or are often polysemous (see Keil, 1986, 1989; Wisniewski & Gentner, 1991; Costello & Keane, 1996). To use an example from this context set, *water* may be a beverage or an ocean, something you drink or something you drown in. In terms of LSA vectors, this leads to a greater variety within the vector and a greater likelihood of chance overlap with a context sentence. In contrast, an *eagle* from the *bird* context set ("The [bird] swooped down…") is much less contextually flexible, and much less likely to be subject to chance overlap with a context sentence vector. This is discussed further in the general discussion.

Related to the beverage disagreements, a pattern exists for those 3 context sets where subjects agreed with the predicted clustering of 9 out of 10 items. One of these was the *bird* set ("Jane liked to listen to…"). Figure 4.7 shows the graph for this clustering, where LSA scores (denoted by filled squares) can be seen falling into two distinct clusters at the bottom-left and top-right corners of the chart. Despite only tenuous initial significance (W=18, $n1$=$n2$=5, p<0.10) subject scores (denoted by open circles) can also be seen to fall into the predicted clusters. The exception to this is the item *peacock*, which was given a relatively high subject score against the context sentence by LSA yet was rated inappropriate by subjects. Removing this item and recalculating the difference between the subject clusters showed significant difference (W=10, $n1$=4, $n2$=5, p<0.05), confirming that without this single-item disagreement, subject scores fell into the predicted clusters. The same effect was achieved by removing *golf* from the *sport* context set ("Sophie was a natural athlete…") and *grapefruit* from the *fruit* context set ("Jimmy liked everything sweet…"), where LSA had predicted the items to be appropriate but subjects disagreed. The previously mentioned pattern emerges with all cases where a single item was in disagreement between subjects and LSA, where LSA had scored it too similar to the context sentence and had placed it in the cluster of appropriate items.

Looking at all cases of incorrect prediction by LSA (a total of 8 out of 70 items), a characteristic of LSA similarity scoring emerges. Of these 8 erroneous predictions, only 1 case existed where LSA predicted an item to be in the inappropriate cluster that subjects then judged appropriate (*tea* in the *beverage* context set). This may have resulted because high and low scores have different traits. High LSA scores (appropriate items) tended to appear on a sliding scale – i.e. often had a relatively large difference of between 0.05 – 0.2 between adjacent items. Low scores (inappropriate items) tended to appear with little variation around the 0.0 point. This has the effect that if an item appears close to the context sentence in semantic space, then LSA may be in error because of an overlap between the item's context vector and the sentence's. Further exposure to the item in a variety of contexts that were truly representative of its usage, would weight its vector so that it moved further from the sentence point and prevent it from being scored as appropriate. However, if an item appears far from the context sentence in semantic space, then LSA is likely to be correct due to absence of overlap in the context vectors. In simple terms, the larger the corpora that the LSA algorithm is run on, the greater the expectation that its representations in semantic space would be contextually accurate. The current incarnation of LSA using corpora for General Reading up to 1st Year College seems to be not large enough.

Experiment 2 – Contextual Typicality vs. Canonical Typicality

Having shown that LSA can be used to predict human judgements of contextual typicality, the question remained whether this typicality in context was significantly different from typicality in a context-free setting. The purpose of this experiment was to test that context was the variable acting in Experiment 1. The hypothesis was that the scores that subjects gave for contextual typicality were indeed influenced by context, and will be different from ratings of canonical typicality. Different subjects were asked to give canonical typicality ratings for the same items and categories used in Experiment 1 these were compared to the contextual ratings. The anticipation was that these two sets of ratings would not correlate significantly.

Method

Subjects

The subjects were 7 native speakers of English and were all volunteers who participated remotely as part of an e-mail questionnaire.

Materials

Materials were as for Experiment 1 except with the category name used instead of the context sentence (7 category sets with 10 possible category members), split into two sections. Each section consisted of 7 category sets with 5 of the possible 10 items. Where there was a pair of category sets that used the same category name, any item was common to both sets was only used once per section. The sections were alternated for each subject. All 35 items within each section were randomly presented for each subject as a pair of [Category Item]. Full tables of materials and scores are available in Appendix D.

Procedure

Subjects read instructions that explained typicality and the 7-point scale as per Rosch (1973) and Armstrong, Gleitman & Gleitman (1983) (See Appendix F).

They were asked not to spend more than 10 seconds deciding on what score to give, and were asked not to change an answer they had earlier given. Any problems or comments could be e-mailed to the experimenter. Steps were taken to ensure that each subject only provided one set of typicality ratings.

Results

Spearman's rank correlation (*rho*) was used to compare the canonical typicality scores to the contextual typicality scores of Experiment 1. The correlation coefficients with significance levels *p* are shown in Table 4.7. Any level of *p* below 0.10 was considered insignificant and marked as such in the table. Reasons for including significance of $p < 0.10$ are the same as those for Simulation 2 and Experiment 1, stemming from small data set size.

Of the 7 category sets, there was no significant rank correlation of canonical and contextual typicality in 5 cases. One category set (*sport*) reached a rank correlation of 0.668, which is significant to $p < 0.05$ when N=10. Figure 4.6 shows a graph for this category set, which despite the correlation has clear differences visible between canonical (denoted by filled squares) and contextual (denoted by open circles) typicality scores. One other set (*beverage*) had a rank correlation of 0.561 (p<0.10) and this category set also displays many differences in score.

Table 4.7: Rank correlations *rho* with significance *p*
for contextual (Experiment 1) and canonical (Experiment 2) typicality scores

Original context sentence	Category	Rank correlation *p*	Significance *p*
Stacy volunteered to milk…	Animal	0.355	insignificant
Fran pleaded with her father…	Animal	0.067	insignificant
The [bird] swooped down…	Bird	0.406	insignificant
Jane liked to listen to…	Bird	0.433	insignificant
Jimmy loved everything sweet…	Fruit	0.539	insignificant
Sophie was a natural athlete…	Sport	0.688	$p < 0.05$
During the mid morning break…	Beverage	0.561	$p < 0.10$

Figure 4.7 shows a graph for the category set *bird* (original context "The [bird] swooped down…"). Contextual typicality scores from Experiment 1 (denoted by open circles) can be seen falling into two clusters – 4 appropriate and 6 inappropriate. This set received 10/10 agreement with LSA predictions. In contrast, the canonical typicality scores from this experiment can be seen scattered throughout the graph, and do not correlate significantly with the contextual ratings (*rho*=0.406).

Discussion

In all 7 category sets, canonical typicality was found to differ from contextual typicality from Experiment 1, this proving the hypothesis correct. Not all category sets achieved clearly insignificant correlation, and it is worth examining one of this that did not – that of *sport*, shown in Figure 4.8. Here, the correlation coefficient between canonical and contextual scores was found to be 0.688 (p<0.05), yet 6 out of the 10 category items show clear changes in typicality judgements. Those that altered the most between context and context-free settings were *football* and *fencing*, judged equally typical at 3.3 in the context given in Experiment 1 ("Sophie was a natural athlete…"). Yet for the canonical category, *football* is judged to be the most typical member at 1.1, while with a score of 5.7 *fencing* is judged the second most atypical member. Smaller differences appear with *golf* (which was judged more canonically typical than its contextual counterpart) and the items *handball*, *cricket* and *darts* (judged more typical in the context sentence than canonically). The remaining 4 items (*basketball*, *tennis*, *hockey* and *bowls*) had very similar canonical and contextual typicality scores. The presence of these differences is important (also found in the significance of $p < 0.05$ and marginal significance of $p < 0.10$ respectively were artifact kinds. This is interesting, as it appears to reinforce the point made earlier that artifact kinds are more contextually flexible as they have a greater variety of functional and relational roles. To link with Rosch's (1978) statement, typicality for artifact kinds is determined as substitutability across a wider variety of contexts than for natural kinds. The net effect of this is that there is a greater chance that typicality

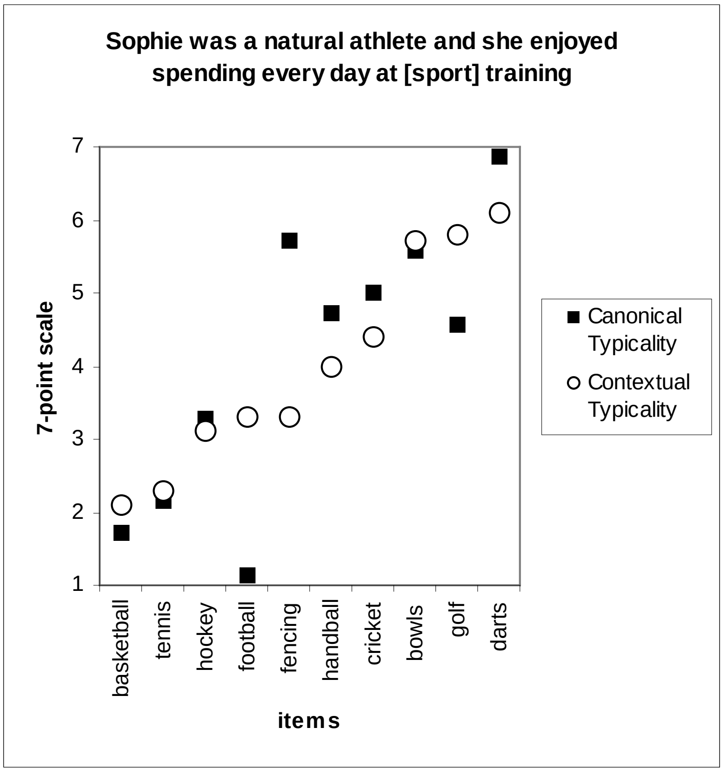

Figure 4.8: graph of subject scores for canonical and contextual typicality of s*port* (note visible differences between canonical and contextual typicality)

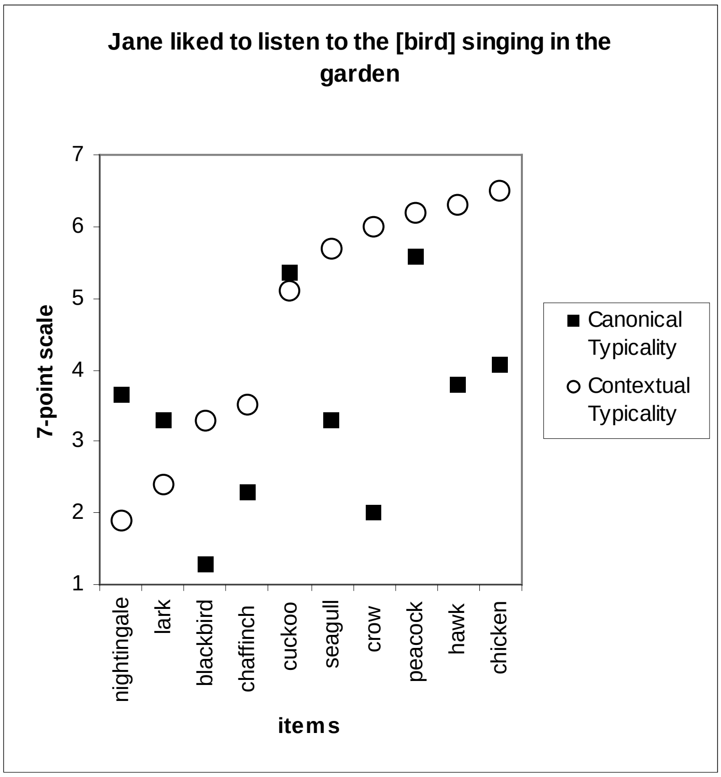

Figure 4.9: graph of subject scores for canonical and contextual typicality of *bird* (note lack of correlation between canonical and contextual typicality)

ratings for an artifact kind item in a given context will correspond to its canonical ratings. Natural kind items, being less contextually flexible in the first place, are more subject to large variations in typicality when a specific context is introduced. On the quantity of data here, this remains speculation but could make for interesting further study.

Chapter 5 General Discussion

Introduction

This thesis began with a review of the literature in the field of categorisation and concepts, before proposing co-occurrence techniques as a possible tool for modelling the empirical data. Simulation work has shown that LSA can separate categories, and succeeds in modelling subject data from a variety of categorisation experiments, such as typicality and context effects. Further experimental work confirmed that context changes the typicality gradients of categories, and that LSA may be used to predict these changes. These results have some interesting repercussions for theories of categorisation and concepts.

Caveats of Co-occurrence Models

Firstly, it is important to note the limitations of using a model such as LSA; some issues are general to all data-intensive techniques and some are specific to LSA itself. It must be noted here that any claims co-occurrence techniques may make with respect to modelling categorisation are necessarily limited to linguistic categorisation. That is, much of human categorisation is grounded in perception, and the only subset which LSA has modelled here is that which is grounded in language.

One of the most common difficulties of statistical natural language processing is proper selection of the corpora. In the case of using a co-occurrence model for categorisation simulation, the accuracy of the co-occurrence – and thus the accuracy of distances in semantic space – depends on both the size and representativeness of the texts used. As discussed earlier, it is clear that LSA does not have truly representative context vectors build up for many words. A case in point in LSA is the penguin problem: when scored for similarity against *bird*, one of the closest category members is *penguin* at 0.63. This should not arise simply because in a representative sample of English, *bird* should be found in many different contexts in which *penguin* is not. A larger corpus may be expected to contain a more representative variety of contexts for *penguin*, which may adjust the *penguin* vector and cancel out this particular quirk. However, low-frequency words tend to be context-specific (Francis & Kucera, 1982), so simply increasing the corpus size is no guarantee of capturing the relevant contexts for a given word. In principle, words with high frequency will always be modelled more effectively than words of low frequency, irrespective of corpus size.

As well as linguistic representativeness, the accuracy of the corpus itself is important. LSA contains misspellings and typographic errors, which means for every error of this type, a proper lexeme was not credited with a co-occurrence score for that context. Example of this include “marriage” appearing misspelled as “marraige” or the error of someone having typed “electric” as “electirc”. While this may not make a significant difference for most lexemes, any word with a low frequency in the corpus will have a noticeably skewed co-occurrence as a result. Connected to this is the issue that co-occurrence techniques are not sensitive to morphological variations. Although one of the interesting achievements of co-occurrence analysis is the extraction of semantic information without any syntactic parsing, from the perspective of this thesis it would have been desirable for the plural and singular form of the word to occupy the same point in semantic space. The disjointedness of singular and plural nouns could have had a significant effect on the results of the simulations and experiments reported in this thesis. For example, comparing *fruit* to the singular *peach* gives a similarity score of 0.26, but comparing *fruit* to the plural form *peaches* causes the score to rise to 0.68. A lemmatised corpus would to some extent solve this problem of morphological variation, and applying the LSA algorithm to create a lemmatised model of co-occurrence seems a plausible goal.

The last matter concerning the underlying corpus is that of its cultural orientation. For the General Reading up to 1st year college semantic space, texts were taken from novels, newspapers, etc., which for the most part were American English. This became a concern when the author ran experiments on British English and Hiberno-English speakers, as a model of language that was representative of American English was not entirely compatible. For example, the *sport* category sets had to avoid mention of *baseball,* which scored a high 0.70 in LSA and was also judged the most typical member (rating 1.2) in studies by Armstrong, Gleitman & Gleitman (1983). However, as this sport is not commonly played outside the US it is unlikely to be rated as typical by non-American subjects. For some instances, the cultural distinctiveness of a term was not a problem such as for the item *football*, likely to be in the context of American football in LSA but interpreted as soccer in Britain. The use of corpora that are representative of one dialect alone is possibly a flawed decision. An unwelcome effect is for alternatives in spelling to be regarded separately, akin to the issue with plural nouns. The American English “color” and British English “colour” only score 0.23 in similarity, due to the much lower frequency of occurrence of “colour”. If a large number of texts from different dialects were used with the LSA algorithm, the greater number of shared contexts would cause their similarity score to rise closer to that expected.

A Context-Based Theory of Categorisation?

LSA’s success in the earlier simulations and experiments has sketched possibilities for a theory of categorisation based in context. Co-occurrence models of language use a type of representation that is learned from the language alone: how certain words co-occur with other gives rise to clues about their semantic meaning. Gleitman (1990) has discussed a similar approach with regards to first language acquisition, where this type of representation can be easily learned from an individual’s response to their linguistic environment, lending a psychologically plausible base to such a theory.

The Basis and Strengths

As a theory of categorisation, a co-occurrence-like approach is first and foremost grounded in the language alone. This obviates explicit hand-coding of category membership or specification of semantic features, making the objective measures of a context-based model more powerful than one that relies on many parameters. The type of representation accessed by categorisation tasks is not meaning explicitly encapsulated within an object representation, but an emergent set of relationships between items in a context-dependent space.

To take an example:

- "Jimmy ate too many sweets and felt sick"
- "Jimmy ate too many sweets and felt nauseous"

Uses such as this in the language are the basis of co-occurrence techniques, as they allow relations to be built between *sick* and *nauseous* because they co-occur with the same words. In the same way, a child may glean the meaning of *nauseous* from the context given by sentence like this, where he or she might similarly have heard the word *sick*. This allows for a mechanism of evolving representations – if the meaning of a word is represented in terms of its use with other words, every piece of text we read has the potential to alter what we think a word means. It also allows for mistakes to be implicitly corrected.

When this paradigm is moved from language alone to the area of categorisation, it begins to sound comparable with theory theories. However, while both theory theories and context-based theories of categorisation share a focus on relational links, this is where their resemblance ends. Context-based categorisation has a built-in, bootstrapped metric of similarity (and therefore relational links), where constraint of the same is the major problem suffered by theory theories. Also, there is no need to posit anything explicit – be it categories, attributes or relations – with a context model, unlike some of its theory-based counterparts (Johnston-Laird, 1983; Lakoff, 1987a, 1987b; etc.). A context-based account does not require the presence of a relational connection for us to know that small birds are more likely to sing than large birds – rather we deduce this because different small birds have more often co-occurred with singing than different large birds. This even offers an explanation for implicit deduction – why we may not be aware of a connection (bird size and birdsong) until we analyse it.

If any existing theory of categorisation bears a resemblance to a context-based account, it is that of ad hoc categories (Barsalou 1983, 1987). Not only consistent with Barsalou's opinion that our representations are unstable, context-based categorisation can deal with ad hoc categories without making special exceptions. As shown in Simulation 4 and Experiment 1, LSA handles a sentence the same way as an individual word – as a single point in high-dimensional space that represents its meaning in terms of other words. In this respect, every case of contextual categorisation is an ad hoc category. An *animal that one may ride upon* is as valid a category as *animal*, and as ad hoc as *things to save if one's home is on fire*, and typicality remains a matter of substitutability in all cases. It is worth noting that "well-defined" categories (Armstrong, Gleitman & Gleitman, 1983) are no exception to this. As shown in Simulations 2 and 3, typicality in some "well-defined" categories arises because they are not well-defined after all (*female*), while in others it is simply a matter of frequency (*odd number* and *even number*). However, although context-based and ad hoc theories may appear close, key differences exist. Context-based categorisation does not require the complex goal construction of Barsalou's theory, nor (like most theory theories) does it assume such a task requires deep conceptual thought.

Natural versus Artifact Kinds

An interesting pattern emerged from the simulations and experiments; that of the difference between LSA's performance with natural and artifact kinds. The difference in processing of the two kinds is a matter of general agreement between different fields of research, from developmental psychology (Keil, 1986) to concept combination (Costello & Keane, 1996). However, in the absence of traditional assumptions such as semantic features or preordained category membership, LSA still showed a difference in performance between natural and artifact kinds.

A context-based theory of categorisation holds that typicality in both canonical and contextual categories is essentially a matter of substitutability. What makes the distinction between natural and artifact kinds is that artifact kinds tend to be more contextually flexible, due to polysemy and/or greater varieties of relational and functional roles (Keil, 1986, 1989; Wisniewski & Gentner, 1991). There is broader contextual substitutability for artifact kinds, which brings about a greater degree of overlap of co-occurrence between one item and the next – i.e. members of an artifact kind tend to share many of their contexts. This results in fewer differences between canonical and contextual typicality in general, because any given context may have already played a part in deciding the canonical typicality gradient for a large number of items, so using it explicitly will not alter their typicality ratings much. In direct contrast, natural kinds are contextually substitutable in a much narrower scope, which brings about a lesser degree of overlap of co-occurrence between items – i.e. members of a natural kind tend to share only some of their contexts. This results in greater differences between canonical and contextual typicality in general, because any given context may have played a part in deciding the canonical typicality gradient for only a small number of items. Therefore using context explicitly will alter their ratings according to whether the context is new to the item or not.

To help clarify this with an example, consider the natural kind *bird*. Members of the *bird* category such as *robin* and *penguin* share some proportion of their contexts – they will both co-occur with references to preening feathers, flapping wings and opening beaks. In general, *robin* is judged far more typical than *penguin* because it occurs in more of these general *bird* contexts. However, they each have a great number of context that they do not share – a *robin* flies, eats worms, hops, and is found in a garden, while a *penguin* swims, eats fish, walks, and is found near ice and sea. Therefore a context involving swimming will already have played a part in the canonical typicality of *penguin*, and so may have little effect on its substitutability. For *robin* however, the context of swimming is new, and because *robin* is not substitutable in this context it will cause its rating to become atypical. Proportionately, *penguin*'s substitutability will now be greater than that of *robin*, and *penguin* becomes the more typical item in this context. It can now be seen how members of an artifact kind would show less alteration in gradient from canonical to contextual typicality if they shared a large number of contexts – there is no proportionate movement if both items remain substitutable.

The reason for difference in LSA's performance in the simulations and experiments follows from this theory. The narrower substitutability of natural kinds meant that it was easier for the corpus used to capture representative use of the category items. There were some exceptions to this – namely *penguin* – but in general it was possible for the corpus to provide a representative variety of contextual uses for each item, and LSA performed close to subject judgements. However, the broad contextual substitutability of artifact kinds meant that the corpus used did not capture full representative use of the category items. An ideal corpus would have been able to provide the full variety of contextual uses for each item. As LSA did not have a representative variety of contexts to hand for all artifact kind items, it performed less effectively and returned scores that were further from subject judgements.

Where Concepts come in

To this point, the discussion has focussed on the process of categorisation rather than the nature of concepts themselves. The reason for this is that a context-based account of categorisation effectively eliminates the reasons why concepts were assumed in the first place – there is no need for a central encapsulation of meaning with attached features, relations etc. if everything is represented in terms of contextual co-occurrence. Indeed, a context-based account also caters for the dynamic and personal nature of conceptualisation, because contextual co-occurrence is a real-time process and is different for every individual. As Kintsch (1998) puts it:

> "A concept depends on an individual's own experiences and can be determined by goals, emotional state, situational and semantic context."

However, there is really only one point to make in relation to concepts per se – that of cognitive parsimony. There is no benefit to having a neat encapsulated canonical concept when the first introduction of context renders it redundant. This is especially underpinned by the fact that most, if not all, of human communication is embedded in context. It also brings into question the entire methodology of examining canonical categories/concepts. Are empirical studies of typicality, feature-naming, relational sensitivity, etc. in canonical categories a measure of human conceptual thought processing, or the forced response of subjects to contrived tasks set in an artificial laboratory environment? Using a co-occurrence model like LSA is essentially a statistical technique, and any information that it extracts that appears to be conceptual in nature is an illusion. It is a statistical bag of words, not a magic bag of tricks. The fact that it has been shown to successfully model a wide variety of psychological effects usually attributed to conceptual processing begs the question of whether any of these phenomena are actually the result of conceptual thought, or merely something much more shallow.

Conclusions

This thesis has shown that a co-occurrence model of language can be used to simulate a range of subject data from the literature, from basic categorisation, to typicality gradients, to the effects of context on category structure. Experimental work has also demonstrated that LSA may be used to predict subject judgements of typicality and appropriateness of items in a given context, and that these judgements vary from canonical typicality.

The conclusions drawn from this are that co-occurrence techniques, as a statistical tool for language, form the basis of an effective model of human categorisation, with a plausible theory attached. However, this does not come without repercussions. If a mathematical algorithm that operates on the co-occurrence of words, and is insensitive to structure and semantics, is able to extract information that can be used to perform tasks previously attributed to conceptual processing, then it raises the question of whether these tasks are conceptual in nature. The ramification for much of the categorisation literature would be that it may have been based on false premises of conceptual thought.

Ideally, the LSA algorithm (or similar co-occurrence technique) should be applied to a larger, culturally diverse set of lemmatised corpora to help establish as representative a semantic space as possible. At this point, further work needs to be done to re-evaluate the empirical methodology and establish the difference between genuinely conceptual thought, and that which may be modelled and is thus the result of shallow task-demands. The assumptions of many cognitive phenomena as conceptual processing may then be open to challenge.

Bibliography

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. Journal of Experimental Psychology: General, 120, 3-19.

Armstrong, S. L., Gleitman, L. R. & Gleitman, H., (1983). What some concepts might not be. Cognition, 13, 263-308.

Battig, W.F., & Montague, W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology Monographs, 80 (3, Pt. 2).

Barsalou, L. (1983). Ad hoc categories. Memory and Cognition 11, 211-227.

Barsalou, L. W., (1987). The instability of graded category structure: implications for the nature of concepts. In U. Neisser (Ed), Concepts and Conceptual Development. Cambridge University Press.

Barsalou, L. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.) Similarity and analogical reasoning .

Barsalou, L. W. and D. R. Sewell (1985). Contrasting the representations of scripts and categories. Journal of Memory and Language, 24: 646--665.

Borges, J. L. (1962). Funes, the Memorious in Ficciones edited by John Sturrock (original publication 1942; English translation, Grove Press, 1962; reprinted. by Alfred A. Knopf/Everyman, 1993).

Burgess, C. & Lund, K., (1997). Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes, 12, 1-34.

Busemeyer, J.R., Dewey, G.I. & Medin, D.L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorization information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 638-648.

Cohen, B. & Murphy, G. L., (1984). Models of concepts. Cognitive Science, 8, 27-58.

Collins, A. M. & Quillian, M. R., (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behaviour, 8, 240-247.

Costello, F. J. & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. In Nineteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Fehr, B. (1988). Prototype analysis of the concepts of love and commitment. Journal of Personality and Social Psychology, 55, 557-579.

Forster, K.I., & Chambers, S. M. (1973). Lexical access and naming time. Journal of Verbal Learning and Verbal Behavior, 12, 627-635.

Francis, W. N. and Kucera, H. (1982). Frequency analysis of English usage: lexicon and grammar. Houghton Mifflin, Boston.

Gleitman, Lila (1990). The structural sources of verb meanings. Language Acquisition, 1, 3-55.

Goodman, N. (1972). Seven strictures on Similarity. In N. Goodman (Ed.), Problems and Projects. New York: The Bobbs-Merrill Co.

Hahn, U. & Chater, N. (1997). Concepts and Similarity in Lamberts, K. and Shanks, D. (Eds.) Knowledge, Concepts and Categories. Cambridge, MA: The MIT Press.

Hampton, J. A., (1982). A demonstration of intransitivity in natural categories. Cognition, 12, 151-164.

Hampton, J.A. (1997a). Psychological representations of concepts. In: M.A.Conway (Ed.) Cognitive models of memory, pp. 81-110. Hove: Psychology Press/Cambridge: MIT Press.

Hampton, J.A. (1997b). Similarity and Categorization. In: M.Ramscar, U.Hahn, E.Cambouropolos, & H.Pain (Eds.) Proceedings of SimCat 1997: An Interdisciplinary Workshop on Similarity and Categorisation, pp. 103-109. Edinburgh: Department of Artificial Intelligence, Edinburgh University.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. Journal of Verbal Learning and Verbal Behavior, 16, 321-338.

Hartley, J. & Homa, D., (1981). Abstraction of stylistic concepts. Journal of Experimental Psychology: Human Learning and Memory, 7, 33-46.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. Cognitive Science, 4, 71-115.

Johnson-Laird, P. N. (1983). Mental models. Cambridge, MA: Harvard University Press.

Katz, J. J. (1972). Semantic Theory. New York: Harper & Row.

Katz, Jerrold J., & Jerry A. Fodor (1963). The structure of a semantic theory, Language 39, 170-210. Reprinted in J. A. Fodor & J. J. Katz, eds. (1964) The Structure of Language, Prentice-Hall, Englewood Cliffs, NJ.

Keil, F.C. (1986). The acquisition of natural kind and artifact terms. In W. Demopoulous & A. Marras (Eds.), Language learning and concept acquisition. Norwood, NJ: Abex

Keil, F.C. (1987). Conceptual Development and Category Structure. In U. Neisser (Ed.), Concepts and Conceptual Development: Ecological and intellectual Factors in Categorization. Cambridge:Cambridge University Press.

Keil, F. C., (1989). Concepts, kinds and conceptual development. Cambridge: MIT Press.

Kintsch, W., (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Komatsu, L., (1992). Recent views of conceptual structure. Psychological Bulletin, 112, 500-526.

Kucera, H. & Francis, W. N. (1967). Computational analysis of present-day American English. Brown University Press, Providence, RI.

Lakoff, G. (1987a). Cognitive models and prototype theory. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization. Cambridge: Cambridge University Press.

Lakoff, G., (1987b). Women, Fire and Dangerous Things. University of Chicago Press.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In Proceedings of the 19th annual meeting of the Cognitive Science Society. Mawhwah, NJ: Erlbaum.

Landauer, T. K. & Dumais, S. T., (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Lund, K., Burgess, C., & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. Proceedings of the Cognitive Science Society. Hillsdale, N.J.: Erlbaum Publishers.

McCloskey, M. & Glucksberg, S., (1978). Natural Categories: Well-defined or fuzzy sets? Memory and Cognition, 6, 462-472.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. Cognitive Psychology, 11, 1-37.

McNamara, T. P., & Sternberg, R. J. (1983). Mental models of word meaning. Journal of Verbal Learning and Verbal Behavior, 22, 449-474.

Malt, B. & Smith, E. (1984). Correlated properties in natural categories. Journal of Verbal Learning and Verbal Behavior, 23, 250-269.

Medin, D. L. (1989). Concepts and Conceptual Structure. American Psychologist, 44, 1469-1481.

Medin, D.L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & Ortony (Eds.). Similarity and Analogical Reasoning. Cambridge, MA: Cambridge University Press.

Medin, D. & Smith, E. (1984). Concepts and concept formation. Annual Review of Psychology, 35, 113-138.

Medin, D. L. & Shoben, E. J., (1988). Context and Structure in Conceptual Combination. Cognitive Psychology, 20, 158-190.

Medin, D.L. & Schaffer, M.M. (1978). Context Theory of Classification Learning. Psychological Review, 85, 207-238.

Medin, D. L. & Schwanenflugel, P. J., (1981). Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.

Medin, D. L, Wattenmaker, W. D. & Hampson, S. E., (1987). Family resemblance, conceptual cohesiveness and category construction. Cognitive Psychology, 19, 242-279.

Michalski, R. S., (1989). Two-tiered concept meaning, inferential matching and conceptual cohesiveness. In S. Vosniadou and A. Ortony (Eds.), Similarity and Analogical Reasoning. New York: Cambridge University Press.

Miller, G. A., & Johnson-Laird, P. N. (1976). Language and perception. Cambridge, MA: Harvard University Press.

Murphy, G. L. & Medin, D. L., (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

Nickerson, R. (1981). Motivated retrieval from archival memory. In G. H. Bower (Ed.) Nebraska symposium of motivation, Vol. 28. Lincoln: University of Nebraska Press.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104- 114.

Nosofsky, R. M., (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Nosofsky, R.M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, & Cognition, 14, 700-708.

Nosofsky, R. M., Clark, S. E., and Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 282-304.

Osherson, D. N. & Smith, E. E., (1981). On the adequacy of prototype theory as a theory of concepts. Cognition, 11, 35-58.

Ramscar, M. J. A. & Hahn, U. (1998). What family resemblances are not: Categorisation and the concept of 'concept'. 20th Annual Conference of the Cognitive Science Society, LEA, pp 865-870.

Ramscar, M.J.A. & Yarlett, D.G. (2000). A high-dimensional model of retrieval in analogy and similarity-based transfer. In Proceedings of the 22nd Annual Meeting of the Cognitive Science Conference.

Reed, S. K. (1972). Pattern recognition and categorization. Cognitive Psychology, 3, 382-407.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadu & A. Ortony (Eds.), Similarity and analogical reasoning. Cambridge: Cambridge University Press.

Robbins, D., Barresi, J., Compton, P., Furst, A., Russo, M. & Smith, M. A. (1978). The genesis and use of exemplar vs. prototype knowledge in abstract category learning. Memory & Cognition, 6, 473-480.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.) Cognitive Development and the Acquisition of Language. New York, Academic Press.

Rosch, E. (1975a). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104, 192-233.

Rosch, E. (1975b). Cognitive reference points. Cognitive Psychology, 7, 532-547.

Rosch, E., (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), cognition and categorization. Hillsdale, N.J.: Erlbaum.

Rosch, E. & Mervis, C. B., (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Rosch, E. (1983). Prototype classification and logical classification: The two systems. In E.K. Scholnick (Ed.), New trends in conceptual representation: Challenges to Piaget's theory? Hillsdale, NJ: Erlbaum.

Roth, E. M. & Shoben, E. J., (1983). The effect of context on the structure of categories. Cognitive Psychology, 15, 346-378.

Rumelhart, D. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce & W. F. Brewer (Eds.) Theoretical issues in reading comprehension. Hillsdale, NJ: Lawerence Erlbaum Associates.

Russel, Bertrand (1946). A History of Western Philosophy. London: George Allen & Unwin.

Smith, E. E. & Medin, D. L., (1981). Categories and Concepts. Cambridge, MA: Harvard University Press.

Smith, E. E., Shoben, E.J., & Rips, L.J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. Psychological Review, 81, 214-241.

Tversky, A., (1977). Features of similarity. Psychological Review, 84, 327-352.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D. & Medin, D. L., (1986). Linear separability and concept learning: Context, relational properties and concept naturalness. Cognitive Psychology, 18, 158-194.

Wattenmaker, W. D., Nakamura, G. N., & Medin, D. L. (1988). Relationships between similarity-based and explanation-based categorization. In D. Hilton (Ed.), Science and natural explanation: Common sense conceptions of causality. NY: New York University Press.

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G. B. Simpson (Ed.), Understanding word and sentence. Amsterdam: Elsevier.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. Cognitive Science, 18, 221- 281.

Wittgenstein, L. 1953. Philosophical investigations. In G. E. M. Anscombe (Trans.). Oxford: Basil Blackwell.

Yarlett, D.G. and Ramscar, M.J.A. (2000). Structure-Mapping theory and lexico-semantic information. In Proceedings of the 22nd Annual Meeting of the Cognitive Science Conference.

Appendix A

Table A1: LSA similarity scores for items against *city* and *country*

Set of Cities	LSA score to *city*	LSA score to *country*	Set of Countries	LSA score to *city*	LSA score to *country*
Atlanta	0.31	0.17	Australia	0.08	0.27
Beijing	0.15	0.13	brazil	0.21	0.33
Berlin	0.10	0.12	Canada	0.12	0.34
Boston	0.19	0.15	Chile	0.18	0.34
Chicago	0.48	0.20	China	0.04	0.21
Dallas	0.29	0.17	England	0.07	0.31
Dublin	0.26	0.10	Finland	-0.02	0.37
London	0.26	0.21	France	0.07	0.22
Madrid	0.22	0.15	Germany	0.06	0.21
Melbourne	0.22	0.27	Hawaii	0.11	0.16
Miami	0.36	0.16	Ireland	0.10	0.26
Montreal	0.16	0.14	Jamaica	0.13	0.16
Moscow	0.18	0.23	Japan	0.05	0.12
Munich	0.05	0.11	Mexico	0.20	0.24
Nashville	0.22	0.37	Norway	0.04	0.31
Paris	0.25	0.18	Poland	0.08	0.29
Portland	0.23	0.18	Russia	0.10	0.25
Rome	0.21	0.05	Scotland	0.06	0.34
Seattle	0.38	0.23	Spain	0.11	0.17
Sydney	0.23	0.35	Sweden	0.07	0.38
Tokyo	0.40	0.09	Wales	0.06	0.28

Table A2: LSA similarity scores for items against *animal* and *body*

Set of Animals	LSA score to *animal*	LSA score to *body*	Set of Body Parts	LSA score to *animal*	LSA score to *body*
bear	0.15	0.04	ankle	0.11	0.23
camel	0.24	0.01	arm	0.07	0.24
cat	0.18	0.05	ear	0.06	0.09
cow	0.19	0.02	eye	0.09	0.09
dog	0.15	0.04	eyelid	0.04	0.07
dolphin	0.12	0.07	face	0.08	0.13
donkey	0.09	0.02	finger	0.05	0.19
elephant	0.57	0.07	foot	0.18	0.19
ferret	0.12	0.01	forehead	0.06	0.10
frog	0.24	0.08	hand	0.09	0.15
horse	0.13	0.02	head	0.17	0.19
kitten	0.15	0.02	heart	0.08	0.32
lion	0.37	0.01	knee	0.07	0.20
mouse	0.14	0.03	leg	0.14	0.24
pony	0.11	0.01	lip	0.04	0.13
puppy	0.16	0.02	nose	0.14	0.22
rat	0.36	0.05	shoulder	0.07	0.16
seal	0.21	0.11	toe	0.18	0.17
squirrel	0.38	0.03	tongue	0.13	0.20
tiger	0.35	0.07	tooth	0.02	0.04
toad	0.19	0.05	waist	0.07	0.26
whale	0.16	0.08	wrist	0.10	0.28

Appendix B

Table B1: Rosch (1973) categories and ratings, with original and scaled LSA scores

Category + item	LSA score	Scaled LSA score	Subject rating
Bird			
robin	0.52	2.68	1.10
eagle	0.80	1.00	1.20
wren	0.40	3.40	1.40
ostrich	0.57	2.38	3.30
chicken	0.31	3.94	3.80
bat	0.20	4.60	5.80
Crime			
murder	0.75	1.00	1.00
stealing	0.41	2.95	1.30
assault	0.41	2.95	1.40
blackmail	0.14	4.50	1.70
embezzlement	0.27	3.75	1.80
vagrancy	0.25	3.87	5.30
Fruit			
apple	0.47	1.00	1.30
pineapple	0.38	2.00	2.30
strawberry	0.33	2.55	2.30
plum	0.27	3.21	2.30
fig	0.02	5.98	4.70
olive	0.23	3.66	6.20
Science			
chemistry	0.64	1.48	1.00
botany	0.66	1.35	1.70
anatomy	0.41	3.07	1.70
geology	0.71	1.00	2.60
sociology	0.44	2.86	4.60
history	0.24	4.24	5.90
Sport			
football	0.76	1.00	1.20
hockey	0.75	1.05	1.80
gymnastics	0.6	1.78	2.60
wrestling	0.48	2.36	3.00
archery	0.24	3.53	3.90
weightlifting	0.07	4.36	4.70
Vegetable			
carrot	0.28	2.28	1.10
asparagus	0.42	1.23	1.30
celery	0.45	1.00	1.70
onion	0.25	2.51	2.70
parsley	0.39	1.45	3.80
pickle	0.26	2.44	4.40
Vehicle			
car	0.47	1.00	1.00
scooter	0.12	4.65	2.50
boat	0.04	5.48	2.70
tricycle	0.01	5.80	3.50
skis	0.05	5.38	5.70
horse	0.00	5.90	5.90

Appendix B – Table B2: Armstrong, Gleitman & Gleitman (1983) categories and ratings, with original and scaled LSA scores

Category + item	LSA score	Scaled LSA score	Subject rating
Female			
mother	0.04	3.57	1.70
sister	0.01	4.27	1.80
ballerina	0.02	4.03	2.00
housewife	0.15	1.00	2.10
actress	0.04	3.57	2.40
hostess	0.01	4.27	2.70
princess	0.01	4.27	3.00
waitress	0.01	4.27	3.20
chairwoman	0.08	2.63	3.40
policewoman	0.02	4.03	3.90
comedienne	0.00	4.50	4.50
Fruit			
orange	0.34	2.49	1.10
apple	0.47	1.00	1.30
cherry	0.43	1.46	1.70
strawberry	0.33	2.61	2.10
plum	0.27	3.30	2.50
pineapple	0.38	2.03	2.70
watermelon	0.21	3.99	2.90
apricot	0.33	2.61	3.00
coconut	0.44	1.34	4.80
fig	0.02	6.17	5.20
olive	0.23	3.76	6.40
Sport			
baseball	0.70	1.37	1.20
football	0.76	1.05	1.40
soccer	0.77	1.00	1.60
hockey	0.75	1.11	1.80
gymnastics	0.60	1.91	2.80
wrestling	0.48	2.54	3.10
fencing	0.22	3.93	3.50
sailing	0.03	4.94	3.80
bowling	0.51	2.38	4.40
hiking	0.18	4.14	4.60
archery	0.24	3.82	4.80
weightlifting	0.07	4.73	5.10
Vegetable			
carrot	0.28	3.65	1.50
peas	0.68	1.00	1.70
spinach	0.58	1.66	1.70
celery	0.45	2.52	2.60
cabbage	0.42	2.72	2.70
asparagus	0.42	2.72	2.70
radish	0.16	4.44	3.10
peppers	0.44	2.59	3.20
onion	0.25	3.85	3.60
pickle	0.26	3.78	4.80
parsley	0.39	2.92	5.00
pumpkin	0.26	3.78	5.50

Appendix B – Table B2: (continued)

Category + item	LSA score	Scaled LSA score	Subject rating
Vehicle			
car	0.47	1.00	1.00
bus	0.24	3.54	1.80
motorcycle	0.34	2.44	2.20
boat	0.04	5.76	3.30
tractor	0.21	3.88	3.70
wagon	0.10	5.09	4.20
scooter	0.12	4.87	4.50
tricycle	0.01	6.09	4.70
horse	0.00	6.20	5.20
sled	-0.03	6.53	5.20
skis	0.05	5.65	5.60
elevator	0.02	5.98	6.20

Appendix B – Table B3: Malt & Smith (1984) categories and ratings, with original and scaled LSA scores

Category + item	LSA score	Scaled LSA score	Subject rating
Bird			
robin	0.52	1.96	1.11
bluebird	0.56	1.61	1.58
seagull	0.47	2.39	1.74
swallow	0.23	4.47	1.84
falcon	0.54	1.78	2.26
mockingbird	0.38	3.17	2.53
starling	0.50	2.13	2.84
owl	0.45	2.56	3.00
vulture	0.04	6.12	3.16
sandpiper	0.08	5.78	3.53
chicken	0.31	3.78	4.05
flamingo	0.44	2.65	4.63
albatross	0.33	3.60	4.68
penguin	0.63	1.00	5.37
bat	0.20	4.73	6.47
Clothing			
shirt	0.41	1.39	1.06
slacks	0.43	1.13	1.06
dress	0.44	1.00	1.25
sweatshirt	0.16	4.66	2.37
coat	0.28	3.09	2.62
underpants	0.12	5.18	2.69
socks	0.28	3.09	3.37
bathrobe	0.04	6.23	3.62
belt	0.22	3.88	3.81
scarf	0.17	4.53	5.19
cape	0.03	6.36	5.62
gloves	0.43	1.13	5.75
necklace	0.15	4.79	6.06
watch	0.14	4.92	6.12
cane	0.15	4.79	6.75
Fruit			
apple	0.47	1.00	1.75
peach	0.26	3.12	2.19
pear	0.28	2.92	2.75
grape	0.21	3.63	2.87
strawberry	0.33	2.41	3.00
lemon	0.34	2.31	3.12
blueberry	0.25	3.22	3.44
watermelon	0.21	3.63	3.94
raisin	0.18	3.93	4.25
fig	0.02	5.55	4.62
coconut	0.44	1.30	4.94
pomegranate	0.13	4.44	5.50
avocado	0.34	2.31	5.62
pumpkin	0.37	2.01	5.69
olive	0.23	3.43	5.75

Appendix B – Table B3: (continued)

Category + item	LSA score	Scaled LSA score	Subject rating
Flowers			
rose	0.25	3.83	1.12
daisy	0.16	4.43	1.56
carnation	0.22	4.03	1.62
violet	0.14	4.57	2.94
poppy	0.18	4.30	3.62
orchid	0.57	1.67	3.62
marigold	0.54	1.88	3.69
tulip	0.67	1.00	3.69
lily	0.23	3.96	4.00
poinsettia	0.25	3.83	4.12
lilac	0.27	3.69	4.50
dandelion	0.55	1.81	4.56
sunflower	0.44	2.55	5.51
Furniture			
sofa	0.44	1.00	1.21
chair	0.4	1.48	1.26
table	0.32	2.43	1.26
desk	0.22	3.63	1.58
dresser	0.34	2.20	1.79
bed	0.26	3.15	1.84
bookcase	0.38	1.72	2.63
piano	0.11	4.95	3.18
footstool	0.20	3.87	3.26
lamp	0.22	3.63	3.48
mirror	0.15	4.47	4.53
cushion	0.18	4.11	4.74
vase	0.25	3.27	5.21
clock	0.05	5.66	5.37
rug	0.32	2.43	5.37
picture	0.12	4.83	5.42
radio	0.05	5.66	5.47
stove	0.35	2.08	5.47
closet	0.38	1.72	6.00
telephone	0.09	5.18	6.26
Trees			
oak	0.74	1.62	2.25
pine	0.84	1.00	2.44
elm	0.68	1.99	2.50
maple	0.75	1.56	2.56
redwood	0.64	2.24	2.87
sequoia	0.54	2.85	3.75
orange	0.27	4.52	3.94
beech	0.69	1.93	3.94
peach	0.24	4.71	4.12
pear	0.22	4.83	4.19
palm	0.49	3.16	4.25
cypress	0.32	4.21	4.37
dogwood	0.27	4.52	4.62
eucalyptus	0.49	3.16	5.06
bamboo	0.45	3.41	6.19

Appendix C

Table C1: Experiment 1 context sets with scaled LSA scores and contextual typicality mean subject ratings per item

Context Sentence	Cluster	Item	Scaled LSA score	Mean subject rating
Stacy volunteered to milk the [animal] whenever she visited the farm	Appropriate	cow	1.0	1.3
		heifer	3.7	4.5
		sheep	4.6	5.1
		goat	5.1	2.8
	Inappropriate	giraffe	6.2	6.9
		bull	6.3	6.9
		camel	6.7	6.5
		dog	6.8	6.8
		rat	6.8	6.7
		bear	7.0	6.6
Fran pleaded with her father to let her ride the [animal]	Appropriate	mare	1.0	2.0
		stallion	2.0	2.6
		pony	2.6	1.2
		horse	2.8	1.5
		mule	4.1	3.1
	Inappropriate	cow	6.3	4.8
		elephant	6.4	4.1
		dog	6.7	6.9
		bear	6.9	5.9
		tiger	7.0	5.5
The [bird] swooped down on the helpless mouse and carried it off	Appropriate	owl	1.0	1.3
		hawk	4.3	2.7
		falcon	4.7	2.4
		eagle	5.0	1.8
	Inappropriate	penguin	6.3	7.0
		cuckoo	6.4	5.7
		chicken	6.4	6.8
		nightingale	6.4	4.9
		vulture	6.8	4.3
		albatross	7.0	4.8
Jane liked to listen to the [bird] singing in the garden	Appropriate	nightingale	1.0	1.9
		lark	1.8	2.4
		peacock	3.4	6.2
		chaffinch	3.4	3.5
		blackbird	4.3	3.3
	Inappropriate	crow	5.5	6.0
		chicken	6.0	6.5
		cuckoo	6.3	5.1
		hawk	6.5	6.3
		seagull	7.0	5.7

Appendix C – Table C1: (continued)

Context Sentence	Cluster	Item	Scaled LSA score	Mean subject rating
Jimmy loved everything sweet and liked to eat a [fruit] with his lunch every day	Appropriate	grapefruit	1.0	5.7
		plum	1.4	3.5
		orange	1.8	4.1
		apple	2.1	3.1
		banana	2.5	3.9
	Inappropriate	raspberry	5.1	5.0
		lemon	5.1	6.8
		grape	5.5	4.4
		lime	5.5	6.0
		fig	7.0	4.8
Sophie was a natural athlete and she enjoyed spending every day at [sport] training	Appropriate	tennis	1.0	2.3
		golf	2.2	5.8
		basketball	2.2	2.1
		hockey	2.2	3.1
		football	2.2	3.3
	Inappropriate	darts	4.9	6.1
		cricket	4.9	4.4
		handball	5.2	4.0
		fencing	6.1	3.3
		bowls	7.0	5.7
During the mid morning break the two secretaries gossiped as they drank the [beverage]	Appropriate	wine	1.0	5.5
		juice	1.8	3.7
		beer	1.8	6.6
		whiskey	2.4	6.0
		coffee	3.0	1.2
	Inappropriate	tea	4.6	1.4
		water	5.2	4.5
		cola	6.4	4.1
		cocoa	6.6	3.5
		saki	7.0	6.1

Appendix D

Table D1: Experiment 2 category sets with canonical typicality mean subject ratings per item

Original context sentence	Category	Item	Mean Subject Rating
Stacy volunteered to milk…	Animal	cow	2.3
		heifer	3.6
		sheep	2.6
		goat	2.6
		giraffe	4.4
		bull	2.4
		camel	3.3
		dog	1.9
		rat	3.1
		bear	5.4
Fran pleaded with her father…	Animal	mare	2.7
		stallion	1.7
		pony	2.7
		horse	2.7
		mule	3.9
		cow	4.0
		elephant	2.1
		dog	3.7
		bear	3.5
		tiger	1.6
The [bird] swooped down…	Bird	owl	4.0
		hawk	3.4
		falcon	4.0
		eagle	3.5
		penguin	5.6
		cuckoo	4.6
		chicken	2.9
		nightingale	3.9
		vulture	2.9
		albatross	6.4
Jane liked to listen to…	Bird	nightingale	3.6
		lark	3.3
		peacock	1.3
		chaffinch	2.3
		blackbird	5.4
		crow	3.3
		chicken	2.0
		cuckoo	5.6
		hawk	3.8
		seagull	4.1

Appendix D – Table D1: (continued)

Original context sentence	Category	Item	Mean Subject Rating
Jimmy loved everything sweet…	Fruit	grapefruit	1.0
		plum	3.1
		orange	1.9
		apple	2.1
		banana	4.1
		raspberry	5.6
		lemon	5.0
		grape	3.3
		lime	4.3
		fig	3.0
Sophie was a natural athlete…	Sport	tennis	1.7
		golf	2.1
		basketball	3.3
		hockey	1.1
		football	5.7
		darts	4.7
		cricket	5.0
		handball	5.6
		fencing	4.6
		bowls	6.9
During the mid morning break…	Beverage	wine	1.4
		juice	1.4
		beer	3.9
		whiskey	2.6
		coffee	2.3
		tea	1.1
		water	2.9
		cola	4.4
		cocoa	5.7
		saki	2.7

Appendix E

Formula E1

Formula used in Simulation 2 for scaling LSA scores to the 7-point typicality scale – suitable when LSA scores fall in the range of [0, +1]. It operates by grounding the highest LSA score of a set at the scale value of 1, and allowing the lower LSA scores to fall proportionately up the 7-point scale. The formula must be applied to every LSA score individually.
Where X is the LSA score one wishes to scale:

$$\text{Scaled LSA score} \quad = \quad \text{Max LSA score} \; - \; \frac{\text{Max LSA score} - 1}{\text{Max LSA score} * X}$$

Formula E2

Formula used in Experiment 1 for scaling LSA scores to the 7-point typicality scale – suitable when LSA scores fall in the range of [-1, +1]. It operates by grounding the midpoint of the LSA score range at the scale value of 4, and allowing the real LSA scores to fall proportionately on either side of the 7-point scale. This effectively grounds the highest and lowest LSA scores of a set at the scale values of 1 and 7 respectively. The formula must be applied to every LSA score individually.
Where X is the LSA score one wishes to scale:

$$\text{Scaled LSA score} \quad = \quad 4 \; - \; \frac{(\text{Midpoint LSA range} - X) \; * 3}{\text{Midpoint LSA range} * \text{Max LSA score}}$$

Note

Formula E1 is not suitable for sets involving negative LSA scores, as the scaled scores may produce ratings > 7. Formula E2 may also be used for LSA score range of [0, +1]. The scaled scores of Formulae E1 and E2 may differ because Formula E1 guarantees a scaled rating of 1 but no set maximum rating, while Formula E2 guarantees scaled ratings of both 1 and 7. This does not affect rank correlation scores or other statistical measures.

Appendix F

Instructions F1

The following were the subject instructions used in Experiment 1:

> This study has to do with what we have in mind when we use words which refer to categories. Take the word red as an example. Imagine a true red. Now imagine an orangish red...imagine a purple red. Although you might still name the orange-red or the purple-red with the term red, they are not as good examples of red (as clear cases of what red refers to) as the clear true red. In short, some reds are redder than others.
>
> Notice that this type of judgement has nothing to do with how well you like the thing: you can like a purple red better than a true red, but still recognise that the colour you see is not a true red. The same is true for other kinds of categories.
>
> In this experiment, you are asked to judge how good an example of a category an instance is a certain context. You may see a sentence like:
>
> "The girl played the GUITAR while the others sang around the campfire"
>
> You are to rate how good an example GUITAR is on a 7-point scale. A score of 1 (one) means that you feel GUITAR is a very good example of the category (musical instruments) in this context. A score of 7 (seven) mean that you feel that GUITAR fits very poorly with your idea or image of an appropriate instrument in the context of a campfire. A score of 4 (four) means that you feel GUITAR fits moderately well, and so on. Use the other numbers of the 7-point scale to indicate intermediate judgements. You will have to type your answer in the box below each sentence and hit enter to move onto the next sentence.
>
> Don't worry about why you feel that something is or isn't a good example of the category in the context. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it.
>
> There are no 'correct' answers, so whatever seems right to you is a valid response. We are interested in your first impressions, so please don't take too much time to think about any one sentence: try to make up your mind quickly, spending less than 10 seconds on each one.
>
> Please e-mail any problems or questions to the experimenter at the address below.

Appendix F – Instructions F2

The following were the subject instructions used in Experiment 2:

> This study has to do with what we have in mind when we use words which refer to categories. Take the word red as an example. Imagine a true red. Now imagine an orangish red...imagine a purple red. Although you might still name the orange-red or the purple-red with the term red, they are not as good examples of red (as clear cases of what red refers to) as the clear true red. In short, some reds are redder than others.
>
> Notice that this type of judgement has nothing to do with how well you like the thing: you can like a purple red better than a true red, but still recognise that the colour you see is not a true red. The same is true for other kinds of categories.
>
> In this experiment, you are asked to judge how good an example of a category an item is. You may see a pair like this:
>
> "Animal DOG"
>
> You are to rate how good an example of Animal that DOG is on a 7-point scale. A score of 1 (one) means that you feel DOG is a very good example of the category Animal. A score of 7 (seven) mean that you feel that DOG fits very poorly with your idea or image of what a good example of Animal is. A score of 4 (four) means that you feel DOG fits moderately well, and so on. Use the other numbers of the 7-point scale to indicate intermediate judgements. You should type your answer below each pair and scroll down to move onto the next sentence.
>
> Don't worry about why you feel that something is or isn't a good example of the category. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it.
>
> There are no ‘correct’ answers, so whatever seems right to you is a valid response. We are interested in your first impressions, so please don't take too much time to think about any one example and please don’t return to change an answer you have already given. Try to make up your mind quickly, spending less than 10 seconds on each one.
>
> Please e-mail any problems or questions to the experimenter as a reply to this message.