Modelling the Acquisition of Natural Language Categories

Trevor Michael Fountain

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2013

Abstract

The ability to reason about categories and category membership is fundamental to human cognition, and as a result a considerable amount of research has explored the acquisition and modelling of categorical structure from a variety of perspectives. These range from feature norming studies involving adult participants (McRae et al. 2005) to long-term infant behavioural studies (Bornstein and Mash 2010) to modelling experiments involving artificial stimuli (Quinn 1987).

In this thesis we focus on the task of *natural language categorisation*, modelling the cognitively plausible acquisition of semantic categories for nouns based on purely linguistic input. Focusing on natural language categories and linguistic input allows us to make use of the tools of distributional semantics to create high-quality representations of meaning in a fully unsupervised fashion, a property not commonly seen in traditional studies of categorisation. We explore how natural language categories can be represented using distributional models of semantics; we construct concept representations for corpora and evaluate their performance against psychological representations based on human-produced features, and show that distributional models can provide a high-quality substitute for equivalent feature representations.

Having shown that corpus-based concept representations can be used to model category structure, we turn our focus to the task of modelling category acquisition and exploring how category structure evolves over time. We identify two key properties necessary for cognitive plausibility in a model of category acquisition, *incrementality* and *non-parametricity*, and construct a pair of models designed around these constraints. Both models are based on a graphical representation of semantics in which a category represents a densely connected subgraph. The first model identifies such subgraphs and uses these to extract a flat organisation of concepts into categories; the second uses a generative approach to identify implicit hierarchical structure and extract an hierarchical category organisation. We compare both models against existing methods of identifying category structure in corpora, and find that they outperform their counterparts on a variety of tasks. Furthermore, the incremental nature of our models allows us to predict the structure of categories during formation and thus to more accurately model category acquisition, a task to which batch-trained exemplar and prototype models are poorly suited.

Acknowledgements

I owe a huge debt to my supervisor, Mirella Lapata, for her constant and invaluable encouragement, and support over the past four years. There is no aspect of this work which has not been improved by her input and criticism. I am tremendously grateful for the support and assistance of a number of my faculty and student colleagues at the University of Edinburgh. Frank Keller provided considerable assistance throughout this thesis; the formation and presentation of the various elicitation studies presented herein benefited enormously from his comments. For innumerable round-table discussions on the nature of language, idiosyncrasies of programming, and quirks of statistics I would also like to thank a few of my student colleagues: John Pate, Mike Lewis, Michal Dziemianko, Moreno Coco, and Desmond Elliott (among many others). I would also like to thank Katrin Erk and Jason Baldridge at The University of Texas at Austin; while they were not directly involved in its creation, this thesis would not exist but for their support and guidance in years past.

Significant portions of this work were published with the assistance of members of the Cognitive Science Society and the North American Chapter of the Association for Computational Linguistics. Reviewers and attendees of CogSci 2010, CogSci 2011 and NAACL 2012 contributed many valuable comments, suggestions, and criticisms. Support, both financial and otherwise, was provided by the Scottish Informatics and Computer Science Alliance.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Trevor Michael Fountain*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

How the mind organises concepts into categories and uses those categories to make predictions is one of the most deeply studied questions in cognitive science. This task of *categorisation* underlies a broad swath of human cognition – most, if not all, cognitive tasks are either some special case of general-purpose categorisation or use knowledge about categories to enable some other task. Categories enable us to extrapolate from our past experiences, allow us to identify stimuli we've not previously encountered, and provide us with a means of predicting the properties of novel stimuli (Smith and Medin 1981). Without categories we might know the objects we had encountered in the past (e.g. our morning coffee) but each time we encountered some new object we would find ourselves at a loss, unable to connect our past experiences with this new item. No matter how many cups of coffee we've consumed in the past, without the ability to form meaningful categories we'd have no way of applying our knowledge of caffeinated beverages and their uses to the hot, milky liquid in front of us. By recognising the mysterious morning cup as a *coffee* (read: as an object belonging to the category of things collectively referred to as coffee), we can make a number of predictions about its properties – it is likely to be both tasty and rejuvenating – and act accordingly. Likewise, we can apply our ability to recognise entirely novel instances of high-level categories (e.g. hot beverages) based on their similarity to familar objects to make predictions about their unobserved properties (e.g. that tea, like coffee, should be served with milk).

Such categories embody much of our knowledge about the world. While the correct identification of a caffeinated beverage may not seem a monumental achievement of human intelligence, the process involved in drawing the inference is surprisingly complex. What we call 'categorisation' is really two (related) tasks: recognising a novel object as an instance of a particular category and using that category to predict unobserved properties. In recognising a novel object we have access to only a handful of observed features, limited by the domains of the senses: e.g. that it has a brown colour, flows like a liquid, and radiates heat. To predict the value of complex, unobserved features – that the object has a pleasant taste or that, after drinking it, we will feel energised – based on only these limited observations is no mean feat.

In this thesis we address the first task, that of identifying the category or categories to which a novel object belongs. While traditional research on categorisation tends to involve either real-world objects (Eimas and Quinn 1994a, Quinn and Eimas 1996) or artificial stimuli (Quinn 1987, Posner and Keele 1968, Bomba and Siqueland 1983), we focus instead on categories acquired from natural language stimuli, i.e. words, a task we refer to as *natural language categorisation*. While this focus is relevant to a number of specific tasks (e.g. infant word learning (Mervis 1987)) restricting ourselves to linguistic categories in this fashion allows us to quickly and flexibly construct representations of word meaning using the tools of distributional semantics. Where traditional models of categorisation are restricted to concepts for which complex feature data is available (e.g. the feature norms of McRae et al. (2005) or Ruts et al. (2004)), extracting semantic representations from corpora allows us to build categorisation models with theoretically unlimited scope, and to specifically explore the impact of the linguistic environment on category formation.

In this introductory chapter we summarise the principle claims set forth in the remainder of the thesis and give an overview of where this work fits within the broader study of categorisation.

1.1 Categorisation

Any attempt to concisely summarise the study of categorisation will be no more successful than, say, a brief overview of physics. The study of how the mind forms and uses categories is so fundamental to human cognition that it has its roots in everything from neuroscience to child psychology – more than enough to fill a good-sized textbook. Nevertheless, this section sets the background for remainder of the thesis through an overview of the relevant context. We begin by discussing three common views on category representation, followed by a brief discussion of category acquisition (which informs Chapters 3 and 4) and natural language categorisation (which informs Chapter 2).

1.1.1 Theories of Categorisation

General models of categorisation tend to be organised into one of a number of theories based on the abstract representations used to define categories. Before we can discuss the contributions of this thesis or the body of work into which it fits, a brief overview of the standard categorisation theories may be beneficial. While most of the work discussed in Chapters 2-4 falls squarely under the exemplar theory, Chapter 2 includes a comparison of exemplar and prototype models for which this overview may be beneficial. At any rate, the division between the these two (relatively) modern theories underlies almost attempts to model categorisation and category acquisition.

Attempts to define an appropriate representation for categories date back to Aristotle's *Categoriae* (n.d.). In this view, the so-called *classical* approach to categorisation, contends that categories can be represented by a list of features which are both necessary (in that an object lacking a requisite feature for a particular category is excluded) and sufficient (in that any object possessing all of the specified features is included). While this provides a suitable representation for certain types of well-defined categories – geometric shapes, mathematical formulae, and biological entities – constructing a list of necessary and sufficient features for the often-fuzzy real-world categories upon which people rely, not to mention ill-defined categories like sports, quickly becomes impossible.

Indeed, the 'necessary and sufficient' approach to category definition was famously skewered by Wittgenstein (1953), who considered the features required to accurately define the concept of games. As an example, consider the obvious features which one might attribute to games: that they involve two or more people, are competitive, and are a leisure activity. Each of these features has an equally obvious counterexample; card solitaire is a non-competitive game involving a single player, while certain casino games can be played professionally. Wittgenstein uses this example (among others) to argue the insufficiency of the classical approach.

For this deficiency and others, the classical view of categories has been almost entirely supplanted by a pair of more recent approaches to the problem of category representation: the *prototype* (Rosch 1973) and *exemplar* (Medin and Schaffer 1978) views. In the prototype view, categories are represented using a single, prototypical instance; this instance, the category's 'prototype', may be a single discrete member which best typifies the category's distinguishing features or an abstraction whose semantic representation somehow entails those of the category's members, e.g. the category label or centroid in which the relevant features of the category (i.e. those features which would have been present in the category's classical representation) are weighted according to their relative importance within the category. Membership in the category is determined by comparing the observed features of a possible member against those of the prototype; if the (weighted) number of matching features surpasses some threshold then we consider the stimuli to be a member of the category.

Significantly, the prototype approach to categorisation overcomes the problematic 'necessary and sufficient' restriction of classical categories. To return to our earlier example, under the classical view we would have difficulty correctly identifying an iced coffee as a kind of coffee, as it lacks the necessary hot temperature feature; under the prototype view, however, it possess enough other, highly-weighted features of the category – it is still brown, still liquid, and still leaves us feeling energised – to be correctly categorised. Furthermore, a prototype model can predict category judgements that, though straightforward for people, prove impossible for a classical model, like intransitivity of category membership (e.g. Big Ben is a clock, and clocks are furniture, but Big Ben is not furniture; see Hampton (1982)) or differences in typicality between members of the same category (e.g. robins and penguins are both birds, but people are significantly more likely to list robin than penguin when asked to name birds; see Barsalou (1985)).

Contrasting the prototype approach are *exemplar models*, in which the notion of a single abstract representation for categories, central to both classical and prototype models, is abandoned entirely. Instead, categories in an exemplar model are represented by a list of previously encountered members, and novel stimuli are judged to be members of a category based on their similarity to other, known members; our cup of morning brew is not a coffee because it possesses some specified features or meets some criteria but simply because it is similar to coffees we have had in the past. Exemplar models can account for the same phenomena that are explained by prototype models; differences in typicality stem from underlying differences in similarity (a penguin is only moderately similar to other birds). Similarly, intransitivity can be handled simply by changing the features used to perform the similarity calculation (e.g. Big Ben is similar in some respects to clocks, but clocks are similar *in a different respect* to furniture).

Of course, it is likely that the cognitive mechanism underlying category representation relies upon neither exemplars nor prototypes, but rather on some interpolation between these extremes. Taken at face value, the exemplar approach suggests that every instance of a category is stored (Reed 1972); other interpretations consider categories in which only the most typical instances are stored (Rosch 1975) or in which many instances are stored to varying degrees of completeness (Komatsu 1992). Alternatively, Vanpaemel and Storms (2008) propose a model in which the complexity of representation varies across categories, with complex categories represented using clusters of exemplars and simpler categories using a single prototype. Their model neatly encompasses both pure exemplar and prototype models, but at significant computational cost (Stukken et al. 2011).

Finally, it is worth mentioning the *knowledge* approach (or, somewhat amusingly, the *theory theory*) to categories, which asserts that categories are formed on the basis of people's general knowledge about the world, rather than on a single idealised member (i.e. a prototype) or an exhaustive list of possible members (i.e. exemplars). This view is perhaps best illustrated by what Barsalou (1985) called *goal-derived categories*: categories that are defined based on how their members fill some externally-determined role. Consider the category of breakfast foods, consisting of concepts like bacon, eggs, or grits. This is quite clearly a category people can and do form, and about which they can make meaningful judgements, yet there is very little similarity between members, making it difficult to account for using an exemplar model, nor is it easy to construct a prototypical representation for possible breakfast foods. Instead, this category seems to be based on people's general knowledge of food and culture – a much more complex abstraction than can be encapsulated by either exemplars or prototypes. Unfortunately, we devote little time to discussions of this approach, as it is diffucult to model the formation of knowledge-based categories without a more complete model of the world and the acquisition of this sort of world knowledge is considerably beyond the scope of this thesis.

1.1.2 Category Acquisition

Investigations into the mechanisms and processes involved in the acquisition of categories are hamstrung by the difficulty of experimentally observing infants' mental models of the world. Necessarily, studies of category formation in early infancy rely on indirect means of assessing infants' familarity or surprise when presented with known or novel stimuli. In most cases infants are first exposed to identifiable members of a category (the 'training' phase) and then presented with a series of items from either the familar category or its contrast (the 'testing' phase). Differences in response between familar and unfamilar items – the order in which items are handled in a sequential touching task (Starkey 1981), or the duration of visual fixation (Eimas et al. 1994) – allow researchers to determine the extent to which the child has generalised his or her experiences with the training exemplars to the category as a whole.

Unfortunately, such studies are often conducted using categories to which infants are likely to have had some prior exposure, e.g. animals (Eimas and Quinn 1994b) or furniture (Behl-Chadha 1996). As a result it is unclear the extent to which infants' performance is based on their existing knowledge of these categories or on whatever discriminating features they may have observed during the (highly structured) training phase of the study. If the former, the training phase might only serve to condition participants to expect the familar category, with differences in response explained better explained by priming or information access cost (Wood et al. 2010). For most experiments involving infants' categorisation of real-world stimuli, it is difficult or impossible to rule out the influence of existing knowledge on category formation (Quinn 2004). Experiments involving adult participants are if anything more problematic, as adult category learners have a wealth of experience and world knowledge which may (perhaps indirectly) influence category formation.

However, it is clear that infants do not require previous exposure to exemplars, even for novel categories, to discriminate between contrasting categories. Bornstein and Mash (2010) describe two classes of categorisation tasks, *experienced-based* and *incremental* (for the latter they use the term *on line*). Experience-based tasks involve distinct training/testing phases, in which participants can apply pre-existing knowledge or experience (arising either from the training phase or simply from their world experience) to the task of disciminating between categories. Incremental tasks, by contrast, are those in which participants can rely only on their immediate perceptions of the stimuli, and must form representations for novel categories while simultaneously applying those categories to discriminate between stimuli.

Exploring the differences in performance between these two tasks, Bornstein and Mash (2010) investigated the effect of previous exposure to a novel category in 3- to 5-month old infants and found insignificant differences in performance in categorising novel exemplars between infants with at-home experience of a category and those first encountering it in the laboratory. Both groups were able to discriminate equally well between the category and its contrast, and neither group demonstrated any significant preference for exemplars of either category. These results contradict the expectation that prior exposure to a category should facilitate discrimination between its exemplars and those of a contrasting category, and suggest that infants' categories are learnt and applied in an incremental fashion.

1.1.3 Natural Language Categorisation

Most experimental work on category modelling and acquisition has revolved around laboratory experiments involving either real-world objects (e.g. children's toys; Starkey 1981), perceptual abstractions (e.g. photographs of animals; Eimas and Quinn 1994b), or abstract, artificial stimuli (e.g. dot patterns or geometric shapes; Posner and Keele 1968 and Bomba and Siqueland 1983, respectively). In most cases researchers' using abstract or artificial stimuli to explore human categorisation would not assert that participants possess a distinct mechanism for distinguishing between categories of (for example) binary strings, but rather that the task invokes a single, global mechanism for learning and applying categories.

Our own approach is no different, in that we treat word meaning as a proxy for conceptual structure (Murphy 2002) and do not suggest that (semantic) categories of words differ significantly from their categories involving their real-world referents. We refer to this task, of organising words into categories based on their semantics, as *natural language categorisation*. While the idea of modelling categories using words as a stand-in for their referents is of course not a new one, explicitly viewing categorisation as the task of organising words into categories based on meaning allows us make use of powerful ideas from artificial intelligence and computational linguistics.

Previous work that could be described as natural language categorisation has a recurring theme: the use of feature norms to construct semantic representation for word meaning. Feature norms are traditionally collected through norming studies, in which participants are presented with a word and asked to generate a number of relevant features for its referent concept (The most notable of these is probably the multi-year project of McRae et al. (2005), which collected and analysed features for a set of 541 common English nouns). The results of such studies can be interesting in their own right, as the frequency and distribution of generated features can provide considerable insight into the nature of participants' categories – but they can also provide material for evaluating prototype and exemplar models. By using feature norms as a proxy for people's mental representations of concepts (i.e. representations based on their perceptual experiences) we can use categorisation models to predict exemplar and prototype effects.

Existing research into natural language categorisation has used such featural representations to explore a wide range of categorisation-related phenomena. Heit and Barsalou (1996) demonstrated their instantiation principle within the context of natural language concepts, Storms et al. (2000) contrasted exemplar and prototype models using a task-based evaluation, and Cree et al. (1999) used feature-based representations to model semantic priming. In all of these models words are used as a proxy for real-world stimuli, and feature norms used as a proxy for people's perceptual experiences of those stimuli. Our approach is to replace feature norms with representations derived from words' context in text corpora, i.e. to use *distributional semantics* to approximate people's perceptual representations of real-world stimuli. While this approach has been criticised as an inaccurate view of how people acquire and use categories – it is clear that both linguistic and perceptual input can be used to learn categories – it has been shown that the information encoded in people's perceptual representations of concepts is often redundantly encoded in their linguistic experience of those concepts (Riordan and Jones 2011). The use of feature norms as a proxy for people's perceptual representation of concepts can itself be problematic, as participants in norming studies tend to under-report or ignore features which do not help to distinguish members of the category but are nevertheless frequently present (McRae et al. 2005). Furthermore, modelling category acquisition using natural language concepts enables us to explore aspects of categorisation which might otherwise prove problematic (e.g. our investigation of incrementality in Chapter 3).

1.2 Central Claims

This thesis addresses the task of natural language categorisation. Where traditional research into category structure and acquisition has relied on either real-world or artificial stimuli using manually-annoted features, we investigate the use of words as stimuli and extract semantic representations from corpora. Following in the tradition of Rational Analysis (Anderson 1991b) we adopt a modelling approach in which we focus on predicting human performance on categorisation tasks rather than positing an underlying mechanism for general category acquisition.

Our overall objective in this thesis is to model the acquisition of natural language categories from linguistic input in a cognitively plausible fashion, but before we can attempt this we first need to determine an appropriate representation for concepts. Our investigations into the suitability of various distributional methods for constructing semantic representations lead to the first claim of this thesis, that corpus statistics, e.g. word co-occurrence, can be used to construct representations of meaning from which useful semantic categories can be learnt. While most models of categorisation rely on manually-produced feature encodings to represent meaning the use of text corpora can provide an attractive alternative to feature norms, which are generally expensive, time-consuming, and difficult to obtain. We demonstrate this claim by constructing alternative representations for a large set of concepts for which high-quality feature norms already exist, using various easily-obtained corpus statistics, e.g. document- or topic-level co-occurrence. We evaluate these representations within a pair of simple exemplar and prototype models, and show that they can provide an acceptable alternative to their featural equivalents.

With the question of how to appropriately construct concept representations for natural language concepts thus settled, we move on to the task of modelling the acquisition of natural language categories. We formulate category acquisition as an *incremental* task, i.e. one in which there are no distinct training and testing phases and the agent learns and applies category information simultanously. While this is similar in nature to the sequential touching task used to investigate category formation in infants (Starkey 1981), the idea of modelling category formation incrementally is rarely used in the context of category *learning* in either infants (Murphy 2002) or adults. Instead, the nature of laboratory-based categorisation studies tends to result in models of categorisation in which some background knowledge or experience (i.e. the training phase) is used to construct mental representations of categories, which are then employed to solve some subsequent task (i.e. the testing phase). Contrary to this trend, we model category acquisition as an explicitly incremental task. We illustrate the incrementality of our models by applying them to the task of predicting interim category structure *during* category acquisition, in a task in which participants are asked to infer category structure based on limited background knowledge. While our initial models learn only a flat, non-overlapping category structure (Chapter 3), we later develop a model capable of predicting complex, hierarchical structure (Chapter 4).

Both our flat and hierarchical models represent concepts and categories as nodes and sub-graphs within a broad graphical structure. Based on our success with these representations, our final claim in this thesis is that such graphical representations can be used to gain greater insight into category acquisition. Our representations are based on the *semantic networks*, in which concepts are represented as nodes within a graph and relations between concepts as edges between their corresponding nodes. The use of this representation is closely intertwined with the previous claim, in that our use of a graphical representation for concepts eases our modelling of category acquisition as an incremental task. We illustrate this claim by constructing incremental models of category acquisition using both graph-based and probabilistic representations and demonstrating that our graph-based approach, in addition to generally outperforming a comparable probabilistic approach, provides considerable flexibility in that it allows our models to bring in external information in an unsupervised fashion. Using a simple semantic network representation, our models can induce complex, hierarchical structures in which the relationship between high-level categories is automatically inferred.

1.3 A Note on Terminology and Typography

For the sake of consistency I have attempted to use the word *concept* throughout the thesis in reference to an observed stimulus; in the context of natural language categorisation – the categorisation of natural language concepts – *concept* and *word* can be used interchangeably. Additionally, I use the term *exemplar* to denote concepts belonging to a specific category. While a category is quite clearly also a concept, I use the word *category* to denote a higher-level organisation of related concepts; e.g. the concepts (or exemplars, or words) *apple*, *orange*, and *pear* belong to the category FRUIT.

Where it is important this distinction be made clear I follow the typographic convention of the previous sentence: *concepts* are written in italics, while CATEGORIES are written using small caps. This convention is especially helpful in the face of categories with hierarchical organisation, in which categories themselves represent concepts organised into higher-level categories, themselves representing concepts at a yet-higher level, ad nauseum.

1.4 Overview of the Thesis

This thesis is divided into two parts: Chapters 2, which explore low-level details of natural language categorisation, and Chapters 3 and 4, each of which presents a cognitively plausible model of natural language category acquisition.

Chapter 2 introduces the concept of natural language categorisation and surveys a number of distributional methods for automatically constructing representations of word meaning from corpora. We compare the relative performance of simple prototype and exemplar models using each of these representations on a set of three standard categorisation tasks: typicality rating, exemplar generation, and category naming. Our results suggest that simple co-occurrence spaces can provide an acceptable approximation to more traditional high-quality (but expensive to obtain) feature norms.

Chapter 3 makes use of the best of the representations developed in the preceeding chapter to perform our first modelling studies of category acquisition, extending the clustering and similarity based approaches explored in the previous chapter to cover novel stimuli and an unspecified number of categories. Our evaluations in Chapter 2 relied on oracle lists for assigning exemplars to categories, limiting us to simple task-based evaluations and making it difficult to perform an effective analysis of the evaluated representations for constructing categories from novel stimuli. Conversely, the models we develop in this chapter make use of similar semantic representations to those of the preceeding chapter but focus explictly on modelling category formation and on the acquisition of semantics for novel stimuli. We identify two key properties required for cognitive plausibility in modelling category acquisition, *incrementality* and *non-parametricity*, and demonstrate how our models capture both of these aspects. We present two models based on very different internal representations, and evaluate both in a series of experiments to reflect adult and child category acquisition in incremental and non-incremental contexts. Our results show that an unsupervised, graph-based model of category acquisition using simple, corpus-derived semantic representations can approximate human performance on numerous categorisation tasks.

Chapter 4 explores the difficulties inherent in constructing categories with more complex internal structure, and presents a model of category acquisition capable of inducing an organisation of flat categories into a full, hierarchical categorisation. A significant drawback of the models we proposed in the preceeding chapter is their inability to learn anything other than a flat organisation of concepts into categories, making them little more than cognitively-informed clustering algorithms. Our challenge therefore becomes the construction of a model which is capable of acquiring more complex (and correspondingly more plausible, for adult learners) category structure while maintaining the key cognitive properties of *incrementality* and *non-parametricity* Our model for doing so is based on the idea of Hierarchical Random Graphs (Clauset et al. 2008), an algorithm for identifying complex hierarchical structure implicit in non-hierarchical graphs (i.e. graphs without an explicit tree structure, root node, or parent-child relationship). As a result, we re-formulate the task of category acquisition as one of graph partitioning, contrasting its presentation in the previous chapter as a clustering task.

We evaluate induced hierarchies using various corpus-based representations against a gold-standard hierarchy (WordNet), and conclude by demonstrating that our model successfully replicates human performance on an incremental task similar to that of Chapter 3.

Chapter 5 concludes by reviewing the claims set out in this introduction, looking at each in light of results from the preceeding chapters. We summarise the central claims and contributions of this thesis, and discuss their relevance within the greater context of research on categorisation. We end the chapter (and this thesis!) by briefly discussing possible avenues for future work, including the integration of perceptual features into our distributional models and the extension of our hierarchical models to cover more complex (e.g. knowledge-based) categories.

1.5 Published Work

Portions of the material discussed in this thesis has been previously published, in particular chapters 2, 3, and 4. Chapter 2 expands on the study of concept and category representations presented in Fountain and Lapata (2010). Chapter 3 is based on research described in Fountain and Lapata (2011), expanding that work to cover additional corpora. The models and experiments presented in Chapter 4 previously appeared in Fountain and Lapata (2012); they are supplemented in the thesis by an experiment assessing human performance on an incremental version of the hierarchy-induction, along with an elicitation study designed to demonstrate the difficulty of that task.

Chapter 2

Representing Exemplars and Categories

Considerable psychological research has shown that people reason about novel objects they encounter by identifying the category to which these objects belong and extrapolating from their past experiences with other members of that category. This task of *categorisation*, or grouping objects into meaningful categories, is a classic problem in the field of cognitive science, one with a history of study dating back to Aristotle (n.d.). This is hardly surprising, as the ability to reason about categories underlies human cognition and is central to a multitude of other tasks, including perception, learning, and the use of language (Smith and Medin 1981, Murphy 2002).

Numerous theories exist as to how humans categorise objects. These theories themselves tend to belong to one of three schools of thought. In the *classical* (or Aristotelian) view categories are defined by a list of “necessary and sufficient” features. For example, the defining features for the concept BACHELOR might be `male`, `single`, and `adult`. Unfortunately, this approach is unable to account for most ordinary usage of categories, as many real-world objects have a somewhat fuzzy definition and don’t fit neatly into well-defined categories (Wittgenstein 1953, Rosch 1978, Smith and Medin 1981).

Prototype theory (Rosch 1973) presents an alternative formulation of this idea, in which categories are defined by an idealized prototypical member possessing the features which are critical to the category. Objects are deemed to be members of the category if they exhibit enough of these features; for example, the characteristic features of FRUIT might include `contains_seeds`, `grows_above_ground`, and `is_edible`. Roughly speaking, prototype theory differs from the classical theory in two ways: members of the category are not required to possess *all* of the features specified in the prototype, and a concept's membership in a category is determined by a probabilistic weighting of those features.

Although prototype theory provides a superior and workable alternative to the classical theory it has been challenged by the *exemplar* approach (Medin and Schaffer 1978). In this view, categories are defined not by a single representation but rather by a list of previously encountered members. Instead of maintaining a single prototype for FRUIT that lists the features typical of fruits, an exemplar model simply stores those instances of fruit to which it has been exposed (e.g. apples, oranges, pears). A new object is grouped into the category if it is sufficiently similar to one or more of the FRUIT instances stored in memory. Sloman et al. (2001) show that both exemplar and prototype models can provide the best explanation for participants' performance on a category naming task. In their studies, perceptual features (e.g. `is_rectangular` or `made_of_glass`) provide the best explanation when paired with an exemplar model, while combined perceptual and functional features (e.g. `used_for_holding_solids`) perform equally well with both exemplar and prototype models. Their results suggest that the question of how to represent category structure (whether by using a collection of exemplars or a single, weighted prototype) is highly dependent on the choice of feature representation.

This chapter focuses on these two related questions: how to represent words and categories in a model of natural language categorisation. First, we investigate methods for constructing feature representations for words based on statistical analysis of large collections of text. We hypothesise that these can provide a viable alternative to manually-produced feature vectors for natural language categorisation. Specifically, we compare categorisation models that represent concepts using manual, human-produced features against those using corpus-derived features produced by Latent Semantic Analysis (LSA, Deerwester et al. 1990), Latent Dirichlet Allocation (LDA, Griffiths et al. 2007c, Blei et al. 2003; a well-known topic model), Dependency Vectors (DV, Padó and Lapata 2007; a semantic space that takes syntactic information into account), and simple co-occurrence counts transformed using pointwise mutual information (PMI). These semantic representations are used as input to two well-established categorisation models, a simple exemplar model similar in construction to Medin and Schaffer's (1978) context model and a prototype model derived thereof. We evaluate the performance of these two models on three adult categorisation tasks — category naming, typicality rating, and exemplar generation — which have been previously modeled exclusively using feature norms (Storms et al. 2000). Our results indicate that meaning representations constructed using simple co-occurrence tend to outperform more sophisticated alternatives whilst lagging behind feature norms by only a small margin. Regardless of the representation employed, we also find that exemplar models consistently outperform their prototype equivalents when using category labels as a stand-in for more complex prototypical representations.

With these representations in hand, we then explore a number of methods for organising concepts into simple, flat categories using automatic clustering techniques. In the experiments described about we rely on gold-standard categories produced by participants in a category-naming experiment, information which is clearly not available to category learners. Because automatically-induced categories are difficult to label, we replace the task-based evaluation with a cluster F-score metric.

2.1 Related Work

In the past much experimental work has tested the predictions of prototype- and exemplar-based theories in laboratory studies involving categorisation and category learning. These experiments tend to use either perceptual stimuli (e.g. images of natural categories (Eimas and Quinn 1994b)) or artificial categories (e.g. patterns of semi-random dots (Posner and Keele 1968) or abstract geometric shapes (Bomba and Siqueland 1983, Quinn 1987)). Models used in these experiments have focused on how categories and stimuli can be represented (Griffiths et al. 2007a, Sanborn et al. 2006) and how best to formalize similarity. The latter plays an important role in both prototype and exemplar models as generalising categories to include new objects depends on correctly identifying previously encountered items.

In this chapter we focus on the less studied problem of modelling the categorisation of *natural language concepts*. In contrast to the numerous studies using perceptual stimuli or artificial categories, there is surprisingly little work on how natural language categories are learnt or used by adult speakers. A few notable exceptions are Heit and Barsalou (1996) who use natural language concepts to illustrate that detailed information about individual exemplars is encoded in humans' category representations, Storms et al. (2000) who evaluate the differences in performance between exemplar and prototype models on a number of natural categorisation tasks, and Voorspoels et al. (2008) who model typicality ratings for natural language concepts. A common assumption underlying this work is that the meaning of the concepts involved in categorisation can be represented by a set of features (also referred to as properties or attributes).

Featural representations like these have played a central role in psychological theories of semantic cognition and knowledge organization and many studies have been conducted to elicit detailed knowledge of features. In a typical norming study participants are given a series of object names and for each object are asked to name all the characteristic properties of that object. Although these feature norms are often interpreted as a useful proxy of the structure of semantic representations, a number of difficulties arise when working with such data (e.g. Sloman and Rips 1998, Zeigenfuse and Lee 2010). For example, the number and types of attributes generated can vary substantially as a function of the amount of time devoted to each object, and there are many degrees of freedom in the way that responses are coded and analyzed. It is not entirely clear how people generate features and whether all of these are important or relevant for representing concepts. Finally, multiple subjects are required to create a representation for each word, which limits elicitation studies to a small number of words and consequently the scope of any computational model based on these feature norms.

Even when the stimuli in question are of an abstract or linguistic nature the features elicited are assumed to be representative of the underlying referents. As an alternative we propose to model the categorisation of linguistic stimuli according to their distribution in corpora. Words whose referents exhibit differing features likely occur in correspondingly different linguistic contexts; our question is whether these differences in usage can provide a substitute for more traditional featural representations. Such distributional representations have been previously shown to accurately model the acquisition of category structure during child language acquisition (Borovsky and Elman 2006).

The idea that words with similar meaning tend to be distributed similarly across contexts is certainly not a novel one. *Semantic space* models, among which Latent Semantic Analysis (LSA, Landauer and Dumais 1997) is perhaps known best, operationalize this idea by capturing word meaning *quantitatively* in terms of simple co-occurrence statistics (between words and paragraphs or documents). While the distributional context employed often varies between models, from word (Erk 2009) or depedency (Padó and Lapata 2007) input to more exotic representations such as hybrid linguistic and perceptual features (Johns and Jones 2011) or models incorporating visual fixation (Chen et al. 2010), semantic space models have been shown to robustly predict performance effects for a variety of tasks involving word meaning (Jones et al. 2011, McRae and Jones 2012). More recently, *topic models* (Griffiths et al. 2007c) have arisen as a more structured representation of word meaning. In contrast to more standard semantic space models where word senses are conflated into a single representation, topic models assume that words observed in a corpus manifest some latent structure – that word meaning can be represented as a probability distribution over a set of topics (corresponding to coarse-grained senses). Each topic is a probability distribution over words whose content is reflected in the words to which it assigns high probability.

2.2 Representing Concepts

We explore four methods for constructing vector space representations of semantic meaning for words, one based on human produced feature norms (Section 2.2.1) and three based on various distributional methods. The distributional methods employed use different levels of granularity to construct semantic representations; all four rely on co-occurrence, but make use of frequency counts at the word (Section 2.2.2), document (Section 2.2.3), topic (Section 2.2.4), and syntactic relation (Section 2.2.5) levels.

2.2.1 Feature Norms

Many behavioral experiments have been conducted to elicit semantic feature norms across languages. One of the largest samples for English is that collected by McRae et al. (2005), who collected feature norms 541 basic-level concepts – e.g. DOG and CHAIR – with features collected in multiple studies taking place over several years. For each concept several annotators were asked to produce a number of relevant features (e.g. `barks`, `has-four-legs`, and `used-for-sitting`). The production frequency of a feature given a particular concept can be viewed as a form of weighting indicating the feature's importance for that concept. A spatial representation of word meaning can be extracted from the norms by constructing a matrix in which each row represents a word and each column a feature for that word. Cells in the matrix correspond to the frequency with which a feature was produced in the context of a given word. An example of such a space is shown in Table 2.1 (a) (vector components represent production frequencies, e.g. 12 participants thought *has-legs* is a feature of TABLE).

(a) Feature Norms

	has_4_legs	used_for_eating	is_a_pet	...
TABLE	12	9	0	...
DOG	14	0	15	...

(b) PMI

	eat	*sit*	*pet*	...
TABLE	1.04	0.73	0.00	...
DOG	0.73	2.18	1.84	...

(c) LSA

	Document 1	Document 2	Document 3	...
TABLE	0.02	0.98	-0.12	...
DOG	0.73	-0.02	0.01	...

(d) LDA

	Topic 1	Topic 2	Topic 3	...
TABLE	0.02	0.73	0.04	...
DOG	0.32	0.01	0.02	...

(e) DV

	subj-of-walk	subj-of-eat	obj-of-clean	...
TABLE	0	3	28	...
DOG	36	48	19	...

Table 2.1: Example semantic representations for *table* and *dog* using feature norms, PMI-transformed co-occurrence (PMI), Latent Semantic Analysis (LSA), Dependency Vectors (DV), and Latent Dirichlet Allocation (LDA). In the feature (a), PMI (b), and DV (e) space values represent co-occurrence counts; in LSA (c) space values are tf-idf scores; in LDA (d) values are probabilities.

2.2.2 Simple Co-occurrence

The most basic corpus-derived semantic space we consider is a matrix in which each row represents a concept (or rather, a word) and each column a possible co-occurring context word. Each entry correponds to the frequency with which the context word appears within a context window of ±5 surrounding the concept. Following standard practice (S and Kaimal 2012), we transform the raw frequency counts using pointwise mutual information (PMI) in order to identify informative co-occurrences. Example representations in this space are shown in Table 2.1 (b) (vector components represent

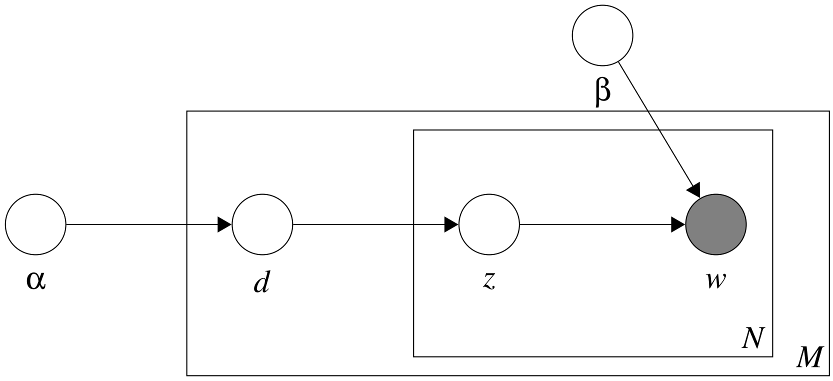


Figure 2.1: Plate notation describing the Latent Dirichlet Allocation model (Griffiths et al. 2007c). d is the distribution of topics within a single document; z is the distribution over observable words w for a topic. α and β function as smoothing parameters for d and w, respectively. M and N indicate the number of documents and the number of observed words in each document, respectively.

PMI-transformed frequency counts).

2.2.3 Latent Semantic Analysis

To create a meaning representation for words LSA constructs a word-document co-occurrence matrix from a large collection of documents. Each row in the matrix represents a word, each column a document, and each entry the frequency with which the word appeared within that document. Because this matrix tends to be quite large it is often transformed via a singular value decomposition (Berry et al. 1995) into three component matrices: a matrix of word vectors, a matrix of document vectors, and a diagonal matrix containing singular values. Re-multiplying these matrices together using only the initial portions of each (corresponding to the use of a lower dimensional spatial representation) produces a tractable approximation to the original matrix. This dimensionality reduction can be thought of as a means of inferring latent structure in distributional data whilst simultaneously making sparse matrices more informative. The resulting lower-dimensional vectors can then be used to represent the meaning of their corresponding words; example representations in LSA space are shown in Table 2.1 (c) (vector components represent tf-idf scores).

2.2.4 Latent Dirichlet Allocation

Where both our PMI and LSA spaces explicitly construct representations of semantic meaning in a vector space based on co-occurrence, word meaning in a Latent Dirichlet Allocation (LDA, Blei et al. 2003) model is expressed as a probability distribution over a set of possible topics. It is based, unlike our previous models, on a generative model for documents: a non-deterministic procedure by which observed documents are supposed to have created. In the generative model each document is viewed as a distribution over a fixed number of topics topics; each topic is itself a distribution over observable words. The individual words in a document are generated by repeatedly sampling first a topic according to the topic distribution of that document and then a single word from that topic. Figure 2.1 shows a graphical representation of the LDA model, in which each observed word (w) is produced by probabilistically sampling a topic z from the document's distribution of topics and an instantiated word w from the z distribution.

Under this framework the problem of meaning representation is expressed as one of statistical inference: given some observed data — words in a corpus, for instance — the model must infer the latent structure from which it was generated. This inference is often accomplished through some form of (e.g. Gibbs) sampling (Griffiths and Steyvers 2004, Phan et al. 2008). Word meaning in LDA is thus represented as a probability distribution over a set of latent topics. To make use of a probabilistic representation in our framework we take the meaning of a word under LDA to be a vector whose dimensions correspond to topics and whose values correspond to the probability of the word given these topics; the likelihood of seeing a word summed over all possible topics is always one. Example representations of words in LDA space appear in Table 2.1 (d) (vector components are topic-word distributions).

2.2.5 Dependency Vectors

Analogously to LSA, the dependency vectors model constructs a co-occurrence matrix in which each row represents a single word; unlike LSA, the columns of the matrix correspond to other words in whose syntactic context the target word appears. These dimensions may be either the context word alone (e.g. `walks`) or the context word paired with the dependency relation in which it occurs (e.g. `subj-of-walks`). Many variants of syntactically aware semantic space models have been proposed in the literature; from these we adopt the framework of Padó and Lapata (2007) where a semantic space is constructed over dependency paths, namely sequences of dependency edges extracted from the dependency parse of a sentence. Three parameters specify the semantic space: (a) the *content selection function* determines which paths contribute towards the representation (e.g. paths of length 1), (b) the *path value function* assigns weights to paths (e.g. it can be used to discount longer paths, or give more weight to paths containing subjects and objects as opposed to determiners or modifiers.), and (c) the *basis mapping function* creates the dimensions of the semantic space by mapping paths that end in the same word to the same dimension. A sample dependency space in shown in Table 2.1 (e) (vector components represent co-occurrence frequencies).

2.3 Representing Categories

The semantic representations described in the preceeding section serve as the input to two categorisation models, representative of the exemplar-based and prototype-based approaches. We derive both of our models from the Context Model (CM, Medin and Schaffer 1978) – while more complex categorisation models could certainly be defined the simplicity of the CM, along with the ease with which it is can be reduced to a simple prototype model, make it ideal for the task of comparing concept representations. In the context model categories are represented by a list of stored exemplars and inclusion of an unknown item in a category is determined by the net similarity between the item and each of the category's exemplars. Specifically, we compute an additive similarity function $\eta_{w,j}$ in which the similarity of a novel item w to a category c is calculated by summing its similarity to all stored items i belonging to c:

$$\eta_{w,c} = \sum_{i \in c} \eta_{w,i} \tag{2.1}$$

where $\eta_{w,i}$ represents the similarity between w and a stored exemplar i belonging to category c. In Medin and Schaffer's original definition of the CM $\eta_{w,i}$ is computed via Manhattan (for discrete features) or Euclidean (for real-valued features) distance, with vectors re-scaled according to category-specific weightings. We modify this somewhat to account for the automatically-derived nature of our concept representations – without supervised training examples it would be difficult to obtain feature weights for the corpus-derived concept representations previously described. To calculate the similarity $\eta_{w,i}$ between a pair of concepts in our exemplar model we compute the cosine of the angle between the unweighted vectors representing w and i:

$$\eta_{w,i} = cos(\theta) = \frac{v_w \cdot v_i}{||v_w||\ ||v_i||} \tag{2.2}$$

Following Vanpaemel and Storms (2008), we can modify Equation 2.1 into a prototype model by replacing the list of stored exemplars with a single ‘prototypical’ instance c_j:

$$\eta_{w,c} = \eta_{w,c_j} \tag{2.3}$$

The similarity between a concept and a category thus reduces to the cosine distance between representations of the concept and the category’s prototype. There is some disagreement over what constitutes an appropriate prototype, with candidates ranging from the single most representative example (e.g. the most ’doglike’ dog that actually exists) to an abstract, idealised representation (e.g. the most ’doglike’ dog that could exist) or a weighted average over all examples (e.g. the average dog over all existing dogs). As with the exemplar model, the lack of feature weightings resulting from the automatically-derived nature of our concept representations rules out the possibility of constructing a prototype as a weighted mean of exemplars; without weighting such a prototype would be overly similar to our exemplar model, simply representing the category as the centroid of its exemplars. Choosing any single exemplar as the prototype presents similar problems, as the prototype model is then essentially an exemplar model with reduced information.

Instead, for the category prototype c_j we use the representation of the category label, e.g. the prototype for the category FRUIT is the semantic representation of the word *fruit*. Substituting category names for prototypes follows naturally from Rosch’s observation that prototypes can substitute for category names (Rosch 1977); similarly, work in textual entailment (specifically, on the subtask of lexical entailment) suggests that distributional representations of high-level concepts (e.g. category names) can be used to make structured predictions about the representations of their hyponyms (Geffet and Dagan 2005, Shnarch et al. 2011).

The similarity function $\eta w, c$ provides the mechanism by which a simple model of categorisation, be it exemplar- or prototype-based, applies knowledge about existing categories to (novel or previously-encountered) stimuli. We can contrast how typical different stimuli are within a common category by comparing their relative similarity to that category, generate likely exemplars by choosing instances with a likelihood relative to their similarity to the category, or predict the category to which a novel stimulus belongs based on its similarity to previously-encountered categories. For the latter task, discussed in greater detail in Section 2.5.1, we implement $\eta_{n,c}$ as a single-best criterion rather than within a Luce choice rule (Luce 1959) for organising concepts into categories. The Luce choice axiom (or Shephard-Luce choice rule) describes the probability of picking a single item i from a possible set c in terms of some absolute weighting w_i. Predicting the correct category for a stimulus w as a Luce choice selection entails picking a category c from the set of possible categories C with probability $P(c)$:

$$P(c) = \frac{\eta_{w,c}}{\sum_{k}^{C} \eta_{w,k}} \tag{2.4}$$

While the Luce choice axiom has been widely used throughout behavioural research (Logan 2003) and enjoys great empirical support (Luce 1977), it is of primary interest as a means of predicting human performance on selection- or attention-allocation tasks, and thus somewhat orthogonal to our goal of contrasting semantic representations. Implementing our prototype and exemplar models using a single-best criterion (i.e. the model always predicts the most likely category for a concept in a category naming tasks) unquestionably limits their cognitive plausibility; crucially, we employ these models only as a means through which we can evaluate various semantic representations for concepts.

2.4 Gold-Standard Category Data

In order to perform any kind of meaningful evaluation of our models we need to obtain or construct a gold-standard dataset reflecting human performance on categorisation tasks. Unfortunately, to the best of our knowledge no such dataset exists for English-language concepts. Storms et al. (2000) extend the feature norming study of Ruts et al. (2004) with manually-produced annotations for a number of categorisation tasks, including category naming, typicality rating, and exemplar generation – but do so for a limited set of only eight categories.

For our dataset we began with the concepts described in McRae et al.'s (2005) feature norming study. Like Ruts et al. (2004), the McRae et al. norms consist of a list of basic-level concepts annotated with descriptive features. For each concept several participants were asked to produce a number of relevant features. The frequency with which a feature was produced for a concept can be viewed as a form of weighting indicating the feature's importance or relevance for that concept; these frequencies can be used to produce a high-quality spatial representation of the concept's meaning. The McRae et al. norms contain 541 basic-level concepts (e.g. *dog* and *chair*).

In this HIT you are given a series of words and asked to label each one with the category to which best belongs. For example, you might assign ”apple” the category ”fruit”, or decide that ”computer” is a member of the category ”device.” Do **not** come up with a single category for entire group – the words are not necessarily related to one another. If you can, try to come up with category labels that are only a single word; for example, don’t use “musical instrument” when “instrument” will do. I have filled in a few examples; you should complete the rest.

	Exemplar	**Category**
EXAMPLE:	pizza	food
EXAMPLE:	calculator	device
	accordion	
	balloon	
	clarinet	
	sailboat	
	lime	
	whale	
	umbrella	
	buffalo	
	dishwasher	
	goldfish	

Figure 2.2: An example category naming task. For each exemplar, participants are asked to generate an appropriate category label.

Unfortunately, the McRae et al. norms do not include any explicit relational information. Because we are interested in using the norms in a model of categorisation it was thus necessary for us to augment concepts with category labels (e.g. *dog* is a kind of *animal*) and typicality ratings. We collected category labels and typicality ratings in a pair of elicitation studies, both conducted using Amazon Mechanical Turk[1], an online labor marketplace which has been used in a wide variety of elicitation studies and has been shown to be an inexpensive, fast, and (reasonably) reliable source of non-expert annotation for simple tasks (Snow et al. 2008).

[1] `http://www.mturk.com`

2.4.1 Collecting Category Names

To obtain category naming information for each of the 541 concepts in the McRae et al. norms we conducted an elicitation study in which each participant was presented with twenty unrelated, randomly selected concepts from McRae et al.'s (2005) data set and asked to label each with the category to which it best belonged. Responses were in the form of free text, i.e. participants were asked to key in a label rather than select one from a list. Each concept was labeled by ten participants; concepts were then grouped according to the resulting categories. Because annotations collected from Mechanical Turk can be noisy we then discarded those categories containing fewer than five unique concepts, leaving 41 categories for 541 exemplars. Figure 2.2 shows an example of the category naming task as presented to participants; the resulting category labels are listed in Table 2.2. The full listing of exemplars organised into their most commonly-labelled category is included in Appendix A.

To fully integrate these labels into the norms (and to enable their use as prototypical representations) it was necessary to collect semantic features for each, in a fashion similar to McRae et al. (2005). To do this, we replicated the norming study of McRae et al. (2005), again using Mechanical Turk. Participants were presented with a single concept (drawn from the set of category labels collected in our previous study) and asked to generate ten relevant features. Instructions and examples were taken from McRae et al. (2005); Figure 2.3 describes the task instructions as presented to participants. For each category label we collected features from 30 participants, resulting in a large number of features per item. These features were then mapped into the features already present in the norms; as in McRae et al. (2005) this mapping was performed manually.

2.4.2 Collecting Typicality Ratings

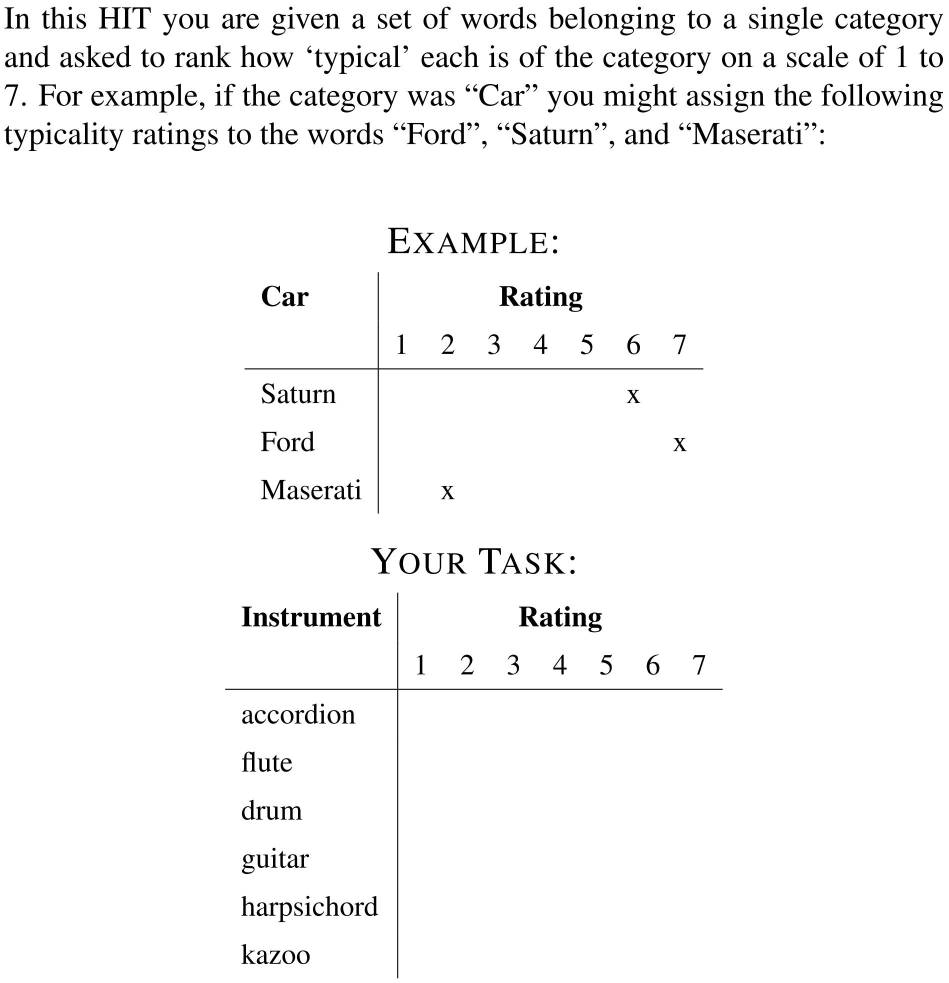
While a mapping between category labels and exemplars is obviously essential to the evaluation of any model of categorisation, typicality ratings have been shown to be indicative of performance on a variety of categorisation tasks (Malt and Smith 1983, Hampton 1993). More significantly, typicality ratings provide a useful proxy for similarity between concepts and categories; in a prototype model highly typical concepts are likely to share a number of features with the prototype. Similarly, in an exemplar model highly typical concepts are likely to be similar to one or more known exemplars of a category.

This experiment is part of an investigation into how people read words for meaning. To help us conduct this work, we need information on what people know about different things in the world. On this task you are given a concept followed by ten blank lines. Please fill in as many of these lines as you can with properties of the concept to which the word refers (one concept per line). Examples of different types of properties would be: physical properties, such as internal and external parts, and how it looks, sounds, smells, feels, or tastes; functional properties, such as what it is used for; where, when and by whom it is used; things that the concept is related to, such as the category that it belongs in; and other facts, such as how it behaves, or where it comes from. Please note that even though many of the words can be thought of as something other than a noun (e.g. camp can refer to the place where your tent is pitched, or the action of camping), all words on the following pages are meant to be considered as nouns only (e.g. camp, the place). Below, we have provided 3 examples to give you an idea of what might be considered a property description of a concept.

duck	cucumber	stove
is a bird	is a vegetable	is an appliance
is an animal	has green skin	produces heat
waddles	has a white inside	has elements
flies	has seeds inside	has an oven
migrates	is cylindrical	made of metal
lays eggs	is long	is hot
quacks	grows in gardens	is electrical
swims	grows on vines	runs on wood
has wings	is edible	runs on gas
has a beak	is crunchy	found in kitchens
has webbed feet	used for making pickles	used for baking
has feathers	eaten in salads	used for cooking food
lives in ponds		
lives in water		
hunted by people		
is edible		

You may be able to think of more and/or different types of properties for these concepts, but these examples should give you an idea of what is requested. Thank you for completing this HIT!

Figure 2.3: Instructions accompaning a feature norming task as presented to participants. These were taken from the feature norming study conducted by McRae et al. (2005); presentation of the instructions was followed by a single concept (e.g. *apple* or *dog*) and a list of ten free-form text entry fields.



In this HIT you are given a set of words belonging to a single category and asked to rank how ‘typical’ each is of the category on a scale of 1 to 7. For example, if the category was “Car” you might assign the following typicality ratings to the words “Ford”, “Saturn”, and “Maserati”:

EXAMPLE:

Car	**Rating**						
	1	2	3	4	5	6	7
Saturn						x	
Ford							x
Maserati		x					

YOUR TASK:

Instrument	**Rating**						
	1	2	3	4	5	6	7
accordion							
flute							
drum							
guitar							
harpsichord							
kazoo							

Figure 2.4: An example typicality rating task. For each exemplar in the given category participants are asked to rate how ‘typical’ that exemplar is among other members of the category.

INSTRUMENT	keyboard	FURNITURE	chair	HOUSING	apartment
REPTILE	rattlesnake	CONTAINER	bin	VEHICLE	bike
CLOTHING	jeans	STRUCTURE	building	VEGETABLE	carrot
HARDWARE	drill	APPLIANCE	stove	BIRD	seagull
HOUSE	cottage	PLANT	vine	TOOLS	hammer
EQUIPMENT	football	UTENSIL	ladle	THING	doll
TOY	surfboard	KITCHEN	dish	RODENT	rat
BUG	beetle	HOME	house	FRUIT	grapefruit
MAMMAL	horse	OBJECT	door	ACCESSORIES	necklace
STORAGE	cabinet	BUILDING	apartment	ANIMAL	cat
DEVICE	stereo	TRANSPORTATION	van	FOOD	bread
GARMENT	coat	FISH	trout	ENCLOSURE	fence
INSECT	grasshopper	SPORTS	helmet	COOKWARE	pan
WEAPON	bazooka				

Table 2.2: Category labels and their most typical exemplar produced by participants in category naming and typicality rating study.

In order to apply our models to the task of predicting typicality ratings it was necessary to conduct an additional elicitation study to obtain gold-standard ratings for our previously-obtained categories. This study was again conducted using Mechanical Turk, and followed a roughly similar setup to the category label elicitation study described in the previous section. Participants were presented with a single category (e.g. FRUIT) along with twenty randomly selected exemplars identified by the previous study as belonging to the category (e.g. *cherry*, *apple*, and *tomato*) and asked to rate the typicality of each listed exemplar among other members of the category. Typicality ratings for each exemplar-category pair were collected from 20 participants and an overall rating for each exemplar was computed by taking their mean. An example of the task as presented to participants is shown in Figure 2.4; Table 2.2 lists the categories obtained in the previous experiment along with the exemplar rated most typical for each.

2.4.3 Evaluating Reliability

We assessed the quality of the category labels and typicality ratings obtained from Mechanical Turk by calculating their *reliability*: the likelihood of a similarly-composed group of participants presented with the same task under the same circumstances pro-

BUILDING	0.61	HARDWARE	0.18	BIRD	0.65
CLOTHING	0.79	ACCESSORIES	0.47	REPTILE	0.91
FOOD	0.69	TOY	0.27	KITCHEN	0.44
COOKWARE	0.83	WEAPON	0.09	PLANT	0.06
HOUSING	0.71	MAMMAL	0.39	DEVICE	0.54
APPLIANCE	0.66	VEGETABLE	0.75	INSTRUMENT	0.37
EQUIPMENT	0.21	THING	0.50	FURNITURE	0.90
HOME	0.74	STRUCTURE	0.54	TOOLS	0.71
UTENSIL	0.25	BUG	0.39	ANIMAL	0.65
GARMENT	0.53	RODENT	0.84	SPORTS	0.19
INSECT	0.01	CLOTHES	0.48	FISH	0.67
CONTAINER	0.47	TRANSPORTATION	0.56	VEHICLE	0.42
ENCLOSURE	0.70	STORAGE	0.20	HOUSE	0.56
OBJECT	0.43	FRUIT	0.56		

Table 2.3: Per-category reliability of human participants on a typicality rating task. Reliability was computed as the split-half overlap and adjusted using the Spearman-Brown prediction forumula.

ducing identical results. We split the collected typicality ratings randomly into two halves and computed the Spearman's ρ correlation between the two; this correlation was averaged across three random splits. These correlations were adjusted by applying the Spearman-Brown prediction formula (Storms et al. 2000, Voorspoels et al. 2008) to compensate for the halving of the test size implicit in computing the split-half correlation. The reliability of the typicality ratings averaged over 41 concepts was 0.51 with a standard deviation of 0.23. The minimum reliability was 0.01 (INSECT); the maximum was 0.91 (REPTILE). Reliability on the category naming task was computed similarly, with an average of 0.67 (standard deviation of 0.15), a maximum of 0.88 (INSTRUMENT), and a minimum of 0.34 (OBJECT).

These reliability figures may seem low compared with Storms et al. (2000) who perform a similar study. However, we note that they conduct a smaller scale experiment; they include only eight common natural language categories (whereas we include 41), and elicit typicality ratings for only 12 exemplars per category (whereas we average $\sim$30 exemplars per category).

Generally speaking, reliability tended to be higher for more specific or familiar categories. On the category naming task (Table 2.4) participants exhibited the highest

BUILDING	0.64	HARDWARE	0.39	BIRD	0.88
CLOTHING	0.80	ACCESSORIES	0.60	REPTILE	0.79
FOOD	0.82	TOY	0.73	KITCHEN	0.55
COOKWARE	0.53	WEAPON	0.76	PLANT	0.58
HOUSING	0.51	MAMMAL	0.83	DEVICE	0.56
APPLIANCE	0.60	VEGETABLE	0.81	INSTRUMENT	0.88
EQUIPMENT	0.60	THING	0.50	FURNITURE	0.78
HOME	0.43	STRUCTURE	0.51	TOOLS	0.52
UTENSIL	0.59	BUG	0.78	ANIMAL	0.82
GARMENT	0.80	RODENT	0.83	SPORTS	0.59
INSECT	0.80	CLOTHES	0.82	FISH	0.80
CONTAINER	0.54	TRANSPORTATION	0.75	VEHICLE	0.74
ENCLOSURE	0.55	STORAGE	0.65	HOUSE	0.59
OBJECT	0.34	FRUIT	0.87		

Table 2.4: Per-category reliability of human participants on a category naming task. Reliability was computed using a split-half correlation and adjusted using the Spearman-Brown prediction formula.

agreement for categories with common, familiar exemplars, e.g. BIRD, (Musical) INSTRUMENT, and FRUIT. Conversely, participants tended to disagree more when asked to name the category for concepts drawn from more generic or loosely-defined categories, e.g. OBJECT or HARDWARE. Category-specific differences on the typicality rating task (Table 2.3) are more difficult to explain; Sloman and Rips (1998) suggest that cultural and linguistic differences between participants as an explanation for similar discrepancies in their experiments, but the online nature of our study prevented us from undertaking any additional investigation of these differences.

2.5 Experiment 1: A Task-Based Comparison of Exemplar and Prototype Models

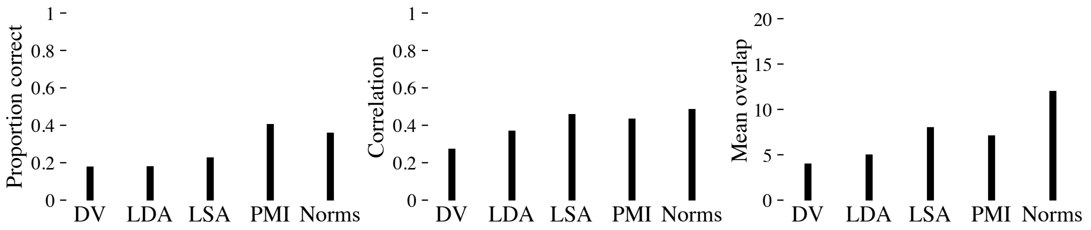
Having collected a set of gold-standard data against which we can assess a model of categorisation we now turn to the task of evaluating our exemplar and prototype models using each of the concept representations described in Section 2.2. Both models were evaluated on three categorisation tasks from Storms et al. (2000): category naming, typicality rating, and exemplar generation.

2.5.1 Categorisation Tasks

In *category naming* the model is presented with a previously unencountered word and must predict the most appropriate category to which it belongs, e.g. the exemplar *apple* would be most correctly identified as a member of the category FRUIT, or (with lesser likelihood) FOOD or TREE. In the exemplar model, we measure the similarity $\eta_{w,c}$ of the novel word against all previously encountered exemplars and select the category with the highest *net* similarity between its exemplars and the word in question; for the prototype model this is the category with the highest similarity between the word and the category's label. Performance on the category naming task was determined in a leave-one-out fashion: a single exemplar was removed from the training examples and then categorised. This was repeated for each exemplar in the training set. The latter consisted of 41 subject-produced category labels each with an average of 30 exemplars.

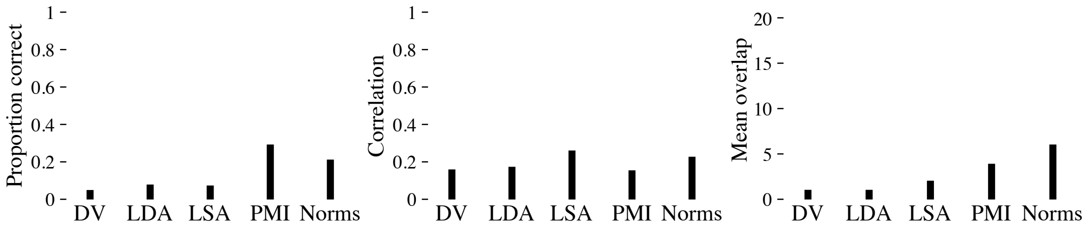
In a *typicality rating task* the model is presented with both an exemplar and label of the category to which it belongs, and must predict the degree to which it is common amongst members of that category. For the category FOOD, for example, *pizza* or *bread* would be considered highly typical exemplars, while *lutefisk* or *black pudding* would likely be considered much more atypical. The predicted typicality rating for a word and a category is simply the similarity between the two. In the exemplar model this is the sum similarity between the word and each of the category's exemplars; in the prototype model this is the similarity between the category's label and the word. The exemplar model was again evaluated in a leave-one-out fashion, with the predicted typicality rating between an exemplar and its gold standard category computed as the similarity between the exemplar and all other gold-standard members of that category, excepting it. Performance on the typicality rating task was evaluated by computing the correlation between the models' predicted typicality ratings and the average value predicted by the participants of our rating study. The dataset included typicality ratings for 1,228 exemplar-category pairs.

In an *exemplar generation* task the model is given a category label and must generate exemplars typical of the category, e.g. for FOOD we might generate *pizza*, *bread*, *chicken*, etc. Given a category the model selects from the exemplars known to belong those that are most typical; typicality is again approximated by word-category similarities as determined by the model-specific $\eta_{w,c}$. As before, the exemplar model was evaluated in a leave-one-out fashion. We evaluate performance on the exemplar generation task by computing the average overlap (across categories) between the ex-



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 2.5: Performance of exemplar model using feature norms and data-driven meaning representations.



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 2.6: Performance of prototype model using feature norms and data-driven meaning representations.

emplars generated by the model and those ranked as most typical of the category by our participants.

2.5.2 Results

Figure 2.5 summarizes our results with the exemplar model and five meaning representations: McRae et al.'s (2005) feature norms (Norms), PMI-transformed word co-occurrence (PMI), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Dependency Vectors (DV). Results are shown for category naming (Figure 2.5a) typicality rating (Figure 2.5b) and exemplar generation (Figure 2.5c). We examined performance differences between models using a χ^2 test (category naming and exemplar generation) and Fisher's r-to-z transformation (to compare correlation coefficients for the typicality rating task).

On category naming the exemplar model performs significantly better with the feature norms than when using DV, LDA, or LSA representations ($p < 0.01$); however, LSA performs significantly better ($p < 0.05$) than DV or LDA and PMI similarly outperforms both feature norms ($p < 0.05$) and all other corpus representations ($p < 0.01$). On typicality rating there is no significant difference between the feature norms, PMI, or LSA. All three are significantly better ($p < 0.01$) than either DV or LDA. Additionally, LDA performs significantly better than DV ($p < 0.05$). On the exemplar generation task the feature norms are significantly better ($p < 0.01$) than any of the corpus-based representations; similarly, both PMI and LSA perform significantly better than LDA or DV ($p < 0.01$) while the difference between LSA and PMI is not significant. LDA again slightly outperforms the dependency space ($p < 0.05$).

Our results with the prototype model are shown in Figure 2.6 and broadly follow a similar pattern. On category naming the feature norms outperform DV, LDA, and LSA ($p < 0.01$); PMI similarly outperforms both the norms and the three other corpus-derived representations ($p < 0.01$). LDA is significantly better than LSA which in turn is better than DV ($p < 0.05$). On typicality rating there is no significant difference between the feature norms and LSA; the difference between LSA and the other three corpus representations is significant ($p < 0.01$). On the exemplar generation task feature norms significantly outperform all other representations ($p < 0.01$); PMI is significantly better than LSA ($p < 0.01$), and LSA is significantly better ($p < 0.01$) than LDA or DV.

2.6 Discussion

In this chapter we have quantitatively evaluated feature norms and alternative corpus-based meaning representations on three natural language categorisation tasks. Perhaps unsurprisingly our results indicate that feature norms are more accurate representations when compared to corpus-based models. As feature norms rely on explicit human judgment, they are able to capture the dimensions of meaning that are psychologically salient; By contrast, corpus-based models learn in an unsupervised fashion and require no human involvement or external knowledge databases such as dictionaries, thesauri or other knowledge repositories.

Overall we find the simple PMI-transformed co-occurrence space to be a reasonable approximation of feature norms, superior to LDA, LSA, and the syntactically more aware dependency vectors. This result is consistent across models (exemplar vs. prototype) and tasks. Importantly, the PMI model (like, indeed, our other corpus-based representations) is language-independent and capable of extracting representations for an arbitrary number of words. By contrast, feature norms tend to cover a few hundred words and involve several subjects over months or years. Albeit in most cases better than our models, feature norms themselves yield relatively low performance on all three tasks we attempted using either an exemplar or prototype model (see Figures 2.5 and 2.6). We believe the reasons for this are twofold. Firstly, McRae et al.'s 2005 norms were not created with categorisation in mind, and we may obtain better predictions with some form of feature weighting (see Storms et al. 2000). Secondly, the tasks seem hard even for humans (as corroborated by our reliability ratings).

The differences in performance between PMI, LSA, LDA, and DV can be explained by differences between the notion of similarity implicit in each. Although LDA and LSA are related – meaning representation in both models is derived from a word-document co-occurrence matrix and the inferred topics in LDA can be viewed as a form of dimensionality reduction – they have distinct notions of similarity. Closely related words in LDA appear in the same *topics*, which are often corpus-specific and difficult to interpret; words belonging to different categories may be deemed similar yet be semantically unrelated. Conversely, our LSA and PMI spaces share the same notion of similarity, but differ in the construction of their respective co-occurrence matrices. The poor performance of the DV model is somewhat disappointing. Our experiments used a large number of dependency relations; it is possible that a more focused semantic space with a few target relations such as coordination and predicate structures (e.g. *Apples and pears are fruits*) may be more appropriate. For these reasons, we rely primarily on concept representations based on the PMI-transformed co-occurrence space for the remainder of the thesis.

Our simulation studies in Experiment 1 suggest that an exemplar model is a better predictor of categorisation performance than a prototype one. This result is in agreement with previous studies (Voorspoels et al. 2008, Storms et al. 2000) showing that exemplar models perform consistently better across a broad range of natural language concepts from different semantic domains. This finding is also in line with studies involving artificial stimuli (e.g. Nosofsky 1992).

Chapter 3

Incremental Models of Category Acquisition

In this chapter we concentrate on the task of acquiring natural language semantic categories and examine how the statistics of the linguistic environment as approximated by large corpora influence category learning. Categories are learnt not only from exposure to the linguistic environment but also from our interaction with the physical world. Perhaps unsurprisingly, words that refer to concrete entities and actions are among the first words being learnt as these are directly observable in the environment (Bornstein et al. 2004). Experimental evidence also shows that children respond to categories on the basis of visual features, e.g. they often generalize object names to new objects on the basis of similarity in shape and texture (Landau et al. 1998, Jones et al. 1991). Nevertheless, we focus on the acquisition of semantic categories from large text corpora based on the hypothesis that simple co-occurrence statistics can be used to capture word meaning quantitatively. The corpus-based approach is attractive for modeling the development of linguistic categories. If simple distributional information really does form the basis of a word's cognitive representation (Harris 1954, Redington and Chater 1997, Braine 1987), this implies that learners are sensitive to the structure of the linguistic environment during language development. As experience with a word accumulates, more information about its contexts of use becomes encoded, with a corresponding increase in the ability of the language learner to use the word appropriately and make inferences about novel words of the same category.

Where our efforts in the previous chapter focused on using text corpora to induce meaning representations for concepts and on using those representations to organise concepts into categories, we now switch to the task of modelling category acquisition. We first describe two properties necessary for a cognitively plausible model of category acquisition, and then develop a pair of models based on differing statistical approaches that satisfy these constraints. We apply these models to a large corpus, both to the original text and to a version enriched with automatically-produced dependency information, as well as to a corpus of child-directed speech. Following this, we assess the performance of both models on a novel categorisation task which highlights their cognitively plausible nature. For the corpus-based experiments we assess performance against a gold-standard set of categories and exemplars; for the cognitive task we compare against categories produced by participants in a human-directed variant of the task.

3.1 Related Work

The task of *categorisation*, in which people cluster stimuli into categories and then use those categories to make inferences about novel stimuli, has long been a core problem within cognitive science. Understanding the mechanisms involved in categorisation, particularly in category acquisition, is essential, as the ability to generalize from experience underlies a variety of common mental tasks, including perception, learning, and the use of language. As a result, category learning has been one of the most extensively studied aspects in human cognition, with computational models that range from strict *prototypes* (categories are represented by a single idealized member which embodies their core properties; e.g. Reed 1972) to full exemplar models (categories are represented by a list of previously encountered members; e.g. Nosofsky 1988) or combinations of the two (e.g. Griffiths et al. 2007a). While the differences between these approaches can have a profound effect on the quality and nature of induced categories, based on our previous exploration of exemplar- and prototype-based models (see Chapter 2 for details), we focus here on a purely exemplar-based approach.

Historically, the stimuli involved in such studies tend to be either concrete objects with an unbounded number of features (e.g. physical objects; Bornstein and Mash 2010) or highly abstract, with a small number of manually specified features (e.g, binary strings, colored shapes; Medin and Schaffer 1978, Kruschke 1993). Most existing models focus on adult categorisation, in which it is assumed that a large number of categories have already been learnt. A notable exception is Anderson's (1991a) rational model of categorisation (see also Griffiths et al. 2007a) where it is assumed that the learner starts without any predefined categories and stimuli are clustered into groups as they are encountered. When a new stimulus is observed, it can either be assigned to one of the pre-existing clusters, or to a new cluster of its own.

This process of learning semantic categories is necessarily incremental. Human language acquisition is bounded by memory and processing limitations, and it is implausible that children process large amounts of linguistic input at once and induce an optimal set of categories. An incremental model learns as it is applied, meaning it does not require separate training and testing phases. Behavioral evidence (Bornstein and Mash 2010), as well as existing models of category acquisition in the context of child-directed speech (Baroni et al. 2007), suggest that this scenario more closely mirrors the process by which infants acquire categories. Having this in mind, we formulate two incremental categorisation models, each differing in the way they represent categories. Both models follow the exemplar tradition — categories are denoted by a list of stored exemplars and inclusion of an unknown item in a category is determined by some notion of similarity between the item and the category exemplars. Previous work (Voorspoels et al. 2008, Storms et al. 2000), as well our own exploration (see Chapter 2) indicates that exemplar models perform consistently better across a broad range of natural language categorisation tasks. This finding is also in line with studies involving artificial stimuli (e.g. Nosofsky 1988). While these studies focus on natural language categories they tend not to specifically address the task of language acquisition; Storms et al. (2000) compare various categorisation models in a natural language context and Voorspoels et al. (2008) use an exempar model to predict typicality ratings for natural language concepts.

In addition to its interest from a cognitive point of view — it cannot be taken for granted that the nature of categorisation performed using artificial categories parallels that involving natural natural language concepts — categorisation of natural language stimuli is fundamental to solving numerous existing problems in natural language processing. Many of these problems can be essentially re-cast as a categorisation task. Word sense discrimination[1] is a good example, a model for clustering words into categories could in many cases double as a word sense discriminator. Other tasks such as the modelling of selectional restrictions (Resnik 1997, Bergsma et al. 2008, Gormley et al. 2011) might benefit from the ability to generalise over categories in a way that does not rely upon expensive, human-produced hierarchies of word relations. Finally,

[1] Word Sense Disambiguation is the task of determining which sense of a word is meant in a particular instance, for words with multiple or ambiguous meaning. It is one of the oldest problems in natural language processing (Weaver 1949), and fundamental to solving a large number of higher-level tasks, e.g. anaphora resolution or discourse analysis.

we hope that the incremental algorithms developed here could be used for large scale data analysis. Standard batch clustering algorithms can quickly become impractical for very large datasets. However, for the algorithms we present here clusters can be updated as data comes along without an expensive re-estimation of the model's parameters.

3.2 Models of Category Acquisition

We present two models of category acquisition, one based on a graph-based representation and another following a probabilistic approach.

Our first model is reminiscent of semantic networks (Collins and Loftus 1975). In this framework, concepts are represented as nodes in a graph and edges represent relationships between such concepts. Although semantic networks are traditionally hand coded by modelers, we learn them from naturally occurring data. In our model, nodes in the graph correspond to words and weighted edges indicate distributional similarity rather than semantic or syntactic relationships. Categories arise naturally in such a representation as densely connected regions or subgraphs. While most research on semantic networks focuses on their use within a larger model of spreading activation (Anderson 1983), they have also been used to gain insight into performance deficits in patients with psychological impairments (Tyler et al. 2000) and to draw comparisons between internet search and memory access (Griffiths et al. 2007b).

Our second model follows a probabilistic approach where categories correspond to topics in a generative model. Topic models have been successful at modeling a wide range of cognitive phenomenal including lexical priming, word association, synonym selection, and reading times (see Griffiths et al. 2007c). In contrast to our use of topic models in the preceeding chapter, in which topic distributions were used as input to a vector-space model of semantics, here we use the topic in which a word is most likely to occur as a proxy for its category. Topics themselves are modeled as probability distributions over words, and can be thought of as a "soft" list of exemplars belonging to the corresponding category. In order to obtain a hard clustering of words into categories we need only compute the topic in which each word is most likely to appear, and assign it to the corresponding category. This model is comparable to our first, semantic network-based model, in that it provides an unsupervised process for inducing categories based on a notion of semantic relatedness; unlike the preceeding model, it performs that process using a purely probabilistic, rather than graph-based, approach.

Algorithm 1: Batch Chinese Whispers

```
initialize;
for $node_i \in Nodes$ do
    class (node) = $i$;
end
while changes do
    for node $\in$ Nodes (in random order) do
        class (target) = class (nearest neighbor)
    end
end
```

Furthermore, both models can be easily adapted to work within cognitive constraints.

Any model of human category acquisition should demonstrate two important features: (1) the input should be processed as it arrives, i.e. the set of clusters is incrementally updated and (2) the set of clusters should not be fixed in advance, but rather determined by the characteristics of the input data. Models obeying the first constraint are henceforth referred to as *incremental*; models obeying the second constraint are said to be *non-parametric*. Note that we use this term to imply only that a model is non-parametric with respect to the number of categories to induce; in our discussions a *non-parametric* model may take other parameters (e.g. smoothing values or a desired number of sampling iterations) but may not require that the number of final categories be specified *a priori*. In the following sections we present a pair of incremental, non-parametric models of category acquisition which employ very different approaches to the task: one is probabilistic, generative model based around the idea of topic models, while the other pairs an algorithm for identifying structure in graphs with the idea of encoding concepts into a semantic network.

3.2.1 Chinese Whispers

The term *semantic network* has been used in many fields to describe a variety of concepts. In Artificial Intelligence and its subfields a semantic network is generally taken to be a directed graph in which nodes represent grounded or un-grounded concepts and edges indicate a directed relationship from one concept to another (Arbib 2002, Cravo and Martins 1993). Following Clauset et al. (2007), we consider a simpler formulation of semantic networks in which a network is composed a graph with edges

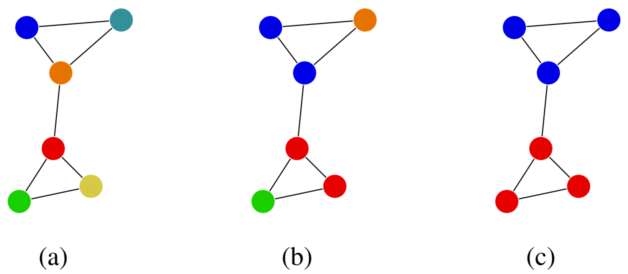


Figure 3.1: An example application of the Chinese Whispers algorithm. The algorithm is initialised by assigning each node in the input graph a novel class (a). On iteration each node in the graph takes on the class of its nearest neighbor (b); after a number of iterations the class assignments stabilise , with the remaining classes corresponding to identifiable subgraphs within the original network (c).

between word nodes. Such a graph is *unipartite*: there is only one type of node, and those nodes can be interconnected freely. While traditional research using semantic networks has focused on performing inference using fully-formed networks to model the organisation of semantic memory, we are argue that they are also well suited to modeling acquisition, as updating the graph to reflect newly acquired information is a straightforward procedure. Here, we propose what is to our knowledge the first graph-based model of category acquisition, in which categories are extracted from a graph representation by identifying well-structured subgraphs within the network.

The task of extracting such subgraphs is generally viewed as a graph clustering problem; Chinese Whispers (CW, Biemann 2006) is one such randomized graph-clustering algorithm that takes as input a graph with weighted edges and produces a hard clustering[2] over the nodes in the graph. It has several desirable properties, including a tendency to converge rapidly and the ability to infer the number of output clusters. The CW algorithm consists of two steps: initialization and iteration. In the initialization step, each node in the graph is assigned a unique class. In the iterative step, each node in the graph (in random order) adopts the highest ranked class in its neighborhood (i.e. the set of nodes with which it shares an edge). Algorithm 1 shows this procedure in pseudocode. CW is in general not guaranteed to converge; in particular, a node with two equally distant nearest neighbors may flip between the classes of

[2] *Hard clustering* is the task of organising a set of items into N discrete clusters in which each item appears in exactly one cluster. In contrast, a items organised into a *soft clustering* may belong to one or more clusters to a varying degree.

Algorithm 2: Incremental Chinese Whispers

```
for $document \in Documents$ do
    for $target,context \in document$ do
        if $target \in graph$ then
            update target representation given context;
            class (target) = class (nearest_neighbour (target));
        else
            add target to graph;
            class (target) = |graph|;
        end
    end
end
```

those neighbors indefinitely. In practice it tends to reach ‘almost-convergence’ quite rapidly (Biemann 2006), which we argue makes it a good fit for modelling the ease and rapidity with which the mind adjusts category structure when presented with new information.

Vanilla CW requires that the entire graph be known before it can be applied, and thus makes no provision for graphs which change over time, as would be expected in an acquisition task. Modifying the CW for use in an incremental setting is straightforward: we need only to update the edges of the graph with newly-encountered input before each iteration step and to run the algorithm until we run out of input to process rather than until convergence (see Algorithm 2).

While applying the incremental CW algorithm to the task of acquiring semantic categories from text, we maintain a weighted, undirected graph in which each node represents a target word and edges between nodes are weighted according to the similarity between words. To compute this similarity, the implementation maintains a running co-occurrence matrix in which each row corresponds to a target word and each column to a possible context word. Similarity between words is computed as the cosine distance between the corresponding rows. Matrix cells are transformed into (positive) pointwise mutual information values (Bullinaria and Levy 2007). Our experiments used a context window centered around a target word (see Chapter 2 for details), however non-symmetric contexts are also possible; target representations are updated according to the context words appearing in the window.

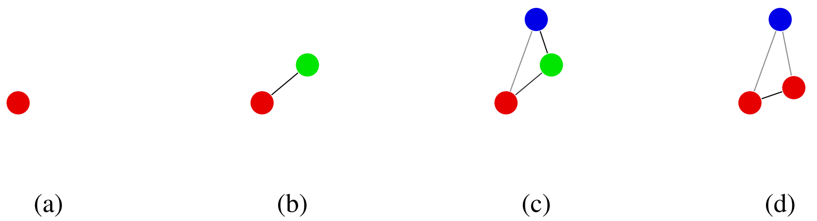


Figure 3.2: An example application of the Incremental Chinese Whispers algorithm. We initialise the algorithm with an empty graph; for the first encountered word (a) we create a new node, assign it the first class, and add it to the graph. For each following encounter of a novel word (b, c) we create a node with a novel class and add it to the graph along with a weighted edges connecting it to any similar nodes. Upon encountering a previously seen word (d) we update its representation, re-compute its edges, and assign it the class of its nearest neighbour.

3.2.2 Topic Model

A great deal of work in recent years has focused on the idea of topic models, in which the meaning of a particular document or word is encapsulated by the latent topics it contains or from which it is generated. Conceptually such models seem appropriate for categorisation tasks, as the notions of “topic” and “category” have much in common.

One particular topic model which has seen wide success is Latent Dirichlet Allocation (LDA, Blei et al. 2003, Griffiths et al. 2007c), which provides a probabilistic model of document generation. In LDA, a document is modeled as a probability distribution over a set of latent topics; similarly, a topic is modeled as a distribution over words. The actual words composing a document are supposed to have been generated by a process of repeatedly sampling first a topic from the document distribution, then a single word from the selected topic. LDA (and generally topic models) can be viewed as a form of a *bipartite* graph consisting of two types of nodes, i.e. words and topics and connections between them.

One drawback to LDA is that it requires the number of topics to be known in advance. As this assumption clearly does not hold in the case of category acquisition, we developed a nonparametric, incremental topic model which is conceptually similar to LDA. This model maintains the generative assumptions of LDA, and much of the same graphical structure; it differs in the addition of a coupling probability (Anderson

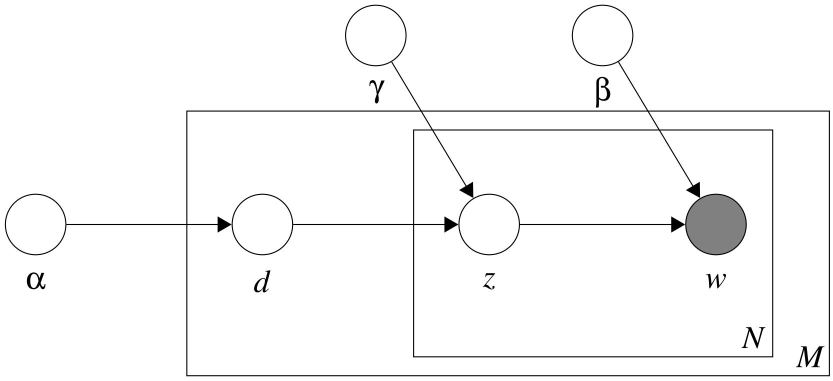


Figure 3.3: A nonparametric topic model which infers the number of topics during training. As in standard LDA, the α and β parameters govern the per-document topic and per-topic word distributions, respectively; this model differs from LDA in the addition of a new parameter γ, which indicates the amount of probability mass reserved for unseen categories (analogous to Anderson's (1990) coupling probability.

1990) used to infer the number of categories during training. Additionally, it performs no final re-estimation of probabilities (as in standard LDA, where re-estimation is performed using Gibbs sampling) in order to maintain incrementality.

In terms of graphical structure our topic model differs from standard LDA (Figure 2.1) by the addition of a third parameter, γ, on the topic distribution. The γ parameter indicates the proportion of probability mass to reserve for a new, previously unseen topic; as additional topics are created the probability of assigning a word to a new topic decreases in relation to γ. α and β act as invisible counts for each topic in a document and each word in a topic, respectively. Combining these parameters with the graphical model in Figure 3.3 yields the following probabilistic model:

$$P(w|z) = \frac{\eta_w^z + \beta}{\sum_x^W (\eta_x^z + \beta)}$$

$$P(z|d) = \frac{(\eta_z^d + \alpha + |W|\beta)(1-\gamma)}{(\sum_y^Z (\eta_y^d + \alpha + |W|\beta)(1-\gamma)) + (\alpha + |W|\beta)\gamma}$$

$$P(z'|d) = \frac{(\alpha + |W|\beta)\gamma}{(\sum_y^Z (\eta_y^d + \alpha + |W|\beta)(1-\gamma)) + (\alpha + |W|\beta)\gamma}$$

$$P(d) = \frac{\sum_y^{Z+z'} (\eta_y^d + \alpha)}{\sum_e^D \sum_y^{Z+z'} (\eta_y^e + \alpha)}$$

where w, z and d represent a word, topic (category), or document, respectively. z' represents a previously unseen topic; a word w assigned to z' is instead assigned to a newly created category initialized to a uniform distribution. The notation η_w^z signifies the number of times word w has appeared in topic z, while η_z^d similarly indicates the count of occurrences of z within document d.

To maintain incrementality, the model performs no re-estimation of probabilities; instead, as each item w of input is encountered it is assigned to a sampled topic z. The relevant document and topic distributions are then updated in accordance with the sampled topic. While these individual predictions are not revised (as in LDA) by subsequent resamplings, predicted topics for subsequent encounters of w change based on the distribution of words and topics; the equations for $P(w|z)$ and $P(z|d)$ are thus analogous to those used during Gibbs sampling in LDA. With additional documents these distributions converge to (hopefully) meaningful topics.

3.3 Evaluating Inferred Categories

In the following section we present three experiments assessing the performance of the CW- and topic-based categorisation models on a category acquisition task. In the first experiment we apply both models to a semantic network induced from a large corpus; in the second we repeat this procedure using a corpus of child-directed speech, with the goal of exploring category acquisition in children. Both of these experiments are conducted in a single batch fashion, entirely ignoring incrementality. We followed these with an experiment to evaluate the models in an incremental context, in which human participants in an elicitation study were asked to produce a series of categorisations in an incremental fashion. We then compare the categorisations induced by both models on this same task to those produced by participants, evaluating how well each model predicts participants' incremental categorisations at each stage of the task. This division between experiments evaluating batch and incremental performance was designed to allow us to disentangle the two key aspects of the models and to assess the methods of exemplar and category representation without introducing additional complexity through an incremental evaluation.

3.3.1 Experiment 2: Category Acquisition From Corpora

Our first goal was to compare our two categorisation models and establish their performance on a large corpus, in order enable a comparison between their differing methods of exemplar and category representation. To do this, we trained both on the British National Corpus (BNC) and compared each model's resulting clustering against a human-produced gold standard. In the following sections we describe how this gold standard was created, discuss how the model parameters were estimated, and explain how the model output was evaluated.

3.3.1.1 Method

As mentioned both models were trained on a version of the BNC which was preprocessed so as to remove stopwords and highly infrequent words, with target words corresponding to frequently-used nouns. The topic model was trained directly on the documents contained in the corpus. It has three free parameters: α (the prior observation count for the number of times a topic is sampled in a document), β (the prior observation count on the number of times words are sampled from a topic), and γ (the probability mass reserved for new topics). For α and β we chose values in accordance with the literature on LDA (Teh et al. 2006); these parameters were set to 1.2 and 0.1, respectively. The γ parameter was tuned on a development corpus (10% of the BNC), with the final value of 0.10. Because of this tuning procedure, all scores reported are from application on the remaining 90% of the BNC not used for development.

Note that the output of the topic model is a set of probability distributions rather than a hard clustering over words. We can nevertheless coerce the model to produce such a clustering by assigning each word to the category (topic) which maximizes its likelihood:

$$category(w) = \underset{z}{\text{argmax}} P(z|w) \tag{3.1}$$

The incremental CW model was trained on noun-centered context windows of ±5, which were extracted from the BNC. As the output of CW is a hard clustering over nodes in the graph, no additional post-processing was required. One obvious question that arises in the context of this experiment is whether using a richer contextual representation yields more accurate categories; we examined this hypothesis by applying the incremental CW algorithm[3] to a dependency-parsed version of the BNC.[4] Specifically, we obtained dependency information from the output of MINIPAR, a broad coverage parser (Lin 2001). To minimize noise this output was restricted to a small set of lexicalized dependency relations: subject, object, and conjunction. The vector space used to compute similarity between words was constructed using context windows, with each word represented as a vector of co-occurrence counts transformed using Pointwise Mutual Information. We deemed this an appropriate space based on the experiments described in Chapter 2, as it was infeasible to apply a more complex representation (e.g. Latent Semantic Analysis) at each step of the incremental algorithm. When applied to the raw BNC dimensions in this space corresponded to possible context words; when applied to the BNC with dependency information dimensions corresponded to lexicalised dependency relations (e.g. OBJ-EAT).

Both models were evaluated based on their clustering of words into semantic categories and their output was compared against similar clusters elicited from human participants. For this evaluation we used the gold-standard category-exemplar mapping produced in Chapter 2; the full list of category names and exemplars is included in Appendix A. This data augments McRae et al.'s (2005) semantic feature norms with category information, and consists of 541 basic-level concepts (e.g. DOG and CHAIR)

[3] Incorporating syntactic information into an incremental topic model is less straightforward, although extensions of the basic LDA model have been proposed that take syntax into account (e.g. Boyd-Graber and Blei 2008).

[4] The use of complete syntactic information such as that obtained through a dependency parse is obviously nonsensical in the context of child category acquisition, as children acquire syntax and semantics simultaneously (Pinker 1994). We argue, however, that it is appropriate for category acquisition in adults, who *do* have access to complex syntactic information and are patently capable of learning new categories based on that information.

REPTILE
salamander, iguana, frog, alligator, rattlesnake, tortoise, crocodile, turtle, toad
FURNITURE
chair, stool, rocker, sofa, cabinet, desk, bookcase, mirror, shelves, bed, drapes, clock, table, bathtub, bureau, cupboard, dresser, fence, cushion, bench, bayonet, armour
FRUIT
peach, yam, nectarine, banana, cantaloupe, apple, plum, raspberry, pear, grape, blueberry, raisin, pineapple, prune, rhubarb, strawberry, lemon, honeydew, orange, tomato, lime, cherry, coconut, olive, grapefruit, tangerine, avocado, pumpkin, cranberry, mandarin

Table 3.1: Example gold standard categories with their exemplars from Fountain and Lapata (2010). The full list of category labels and exemplars is included in Appendix A.

with features collected in multiple studies over several years. The category naming study described in Chapter 2 obtained category labels for 517 of these concepts; in it, participants were presented with a number of nouns chosen at random from the McRae et al. norms and asked to name the category to which each noun belonged. Participant responses were freeform strings, i.e. participants were not provided with a list of possible categories. After adjusting for differences in spelling and conflating synonyms, these responses were used to determine the most "correct" category label for each noun.

Because the norms were originally drawn from a limited number of concepts many of the nouns were labeled with the same category label; we exploited this overlap in order to construct a clustering over the McRae et al. norms in which each cluster corresponds to a subset of nouns assigned the same category label in Chapter 2. Overall, we obtained 32 categories averaging approximately 16 nouns apiece. Examples of the clusters used in our experiments are shown in Table 3.1.

Each model produced a clustering over the nouns taken from the McRae et al. norms which we compared against the human-produced gold standard clustering described above; to evaluate cluster quality we computed the F-score measure described in Agirre and Soroa (2007). Under their evaluation scheme, the gold standard is partitioned into a test and training corpus, the latter of which is used to derive a mapping of

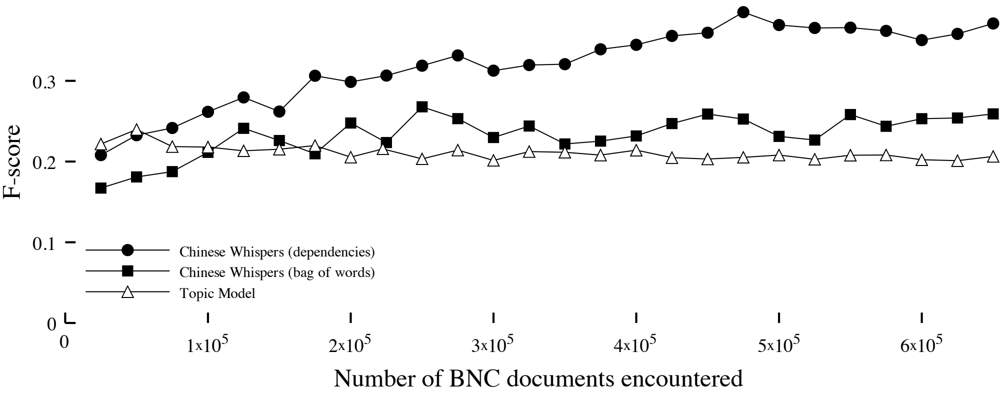


Figure 3.4: Performance of the topic model and Chinese Whispers using dependencies and a bag of words context window.

the induced clusters to the gold standard labels. This mapping is then used to calculate the system's F-score on the test corpus. We calculated F-score as the harmonic mean of precision and recall defined as the number of correct members of a cluster divided by the number of items in the cluster and the number of items in the gold-standard class, respectively (See Chapter 2 for details).

3.3.1.2 Results

All scores were computed according to the F-score measure for unsupervised evaluation described in Agirre and Soroa (2007). Precision and recall were defined as the number of correct members of a cluster divided by the number of items in the cluster and the number of items in the gold-standard class, respectively.

CW and the topic model produced clusters for 517 nouns. As both models are nonparametric, they induce the number of clusters (i.e. categories) from the data as well as which nouns belong to these clusters. The topic model partitioned the target nouns into 167 clusters and CW into 35.

Compared to the gold-standard clustering, the topic model achieved an F-score of 0.179; CW obtained an F-score of 0.212 when using a bag of words context window. The model's performance improved to an F-score of 0.371 when dependency relations were used. To put these numbers into perspective, we also implemented a baseline algorithm that groups nouns into clusters randomly, which achieved an inferior F-score of 0.135. Overall, our results indicate that more fine-grained linguistic information beyond simple co-occurrence is beneficial for categorisation. Figure 3.4 shows how performance on the category acquisition task varies over time (i.e. over the course of encountering all documents in the training set). As can be seen, the quality of clusters produced by CW increases with additional data, i.e. the algorithm's performance improves with more iterations.

3.3.2 Experiment 3: Child Category Acquisition

The primary goal of the preceding experiment was to explore how effectively the two models capture large-scale category information. Of greater interest, however, is modeling children's performance on an acquisition task — determining whether the linguistic input to which children are exposed enables their learning of high-level semantic categories such as those seen in Experiment 1. To answer this question we applied our incremental models to a corpus of child-directed speech and evaluated the resulting categories against the gold-standard clusters used previously.

Intuitively, one would expect the information content and complexity of child-directed speech to increase in relation to the age of the target child. As a result, it should be possible to extract richer categories from speech directed to older children.

3.3.2.1 Method

The CHILDES (MacWhinney 2000) corpus of infant- and child-directed speech was used to construct training documents for both models. CHILDES consists of a large number of transcripts in a multitude of languages, each recording a free-form interactive session between a child and one or more adults (parents); from these the transcripts involving American English speakers (4392 transcripts involving 43 children) were selected, with each grouped according to the child's age in months. All utterances produced by the child were excluded from the final documents, leaving a corpus of child-directed speech organized by target age. While it would be ideal to use a large corpus directed at at single child, so as to accurately capture the linguistic environment in which category learning occurs (Baroni et al. 2007), to our knowledge no such corpus exists.

Both the incremental CW and nonparametric topic model were applied to this corpus using the same parameters as in Section 3.3.1. For CW we used only a bag-of-words representation of context, dropping the dependency information previously employed. While the final result of the experiment in Section 3.3.1 suggests that richer representations, such as those derived from a dependency parse, yield more accurate

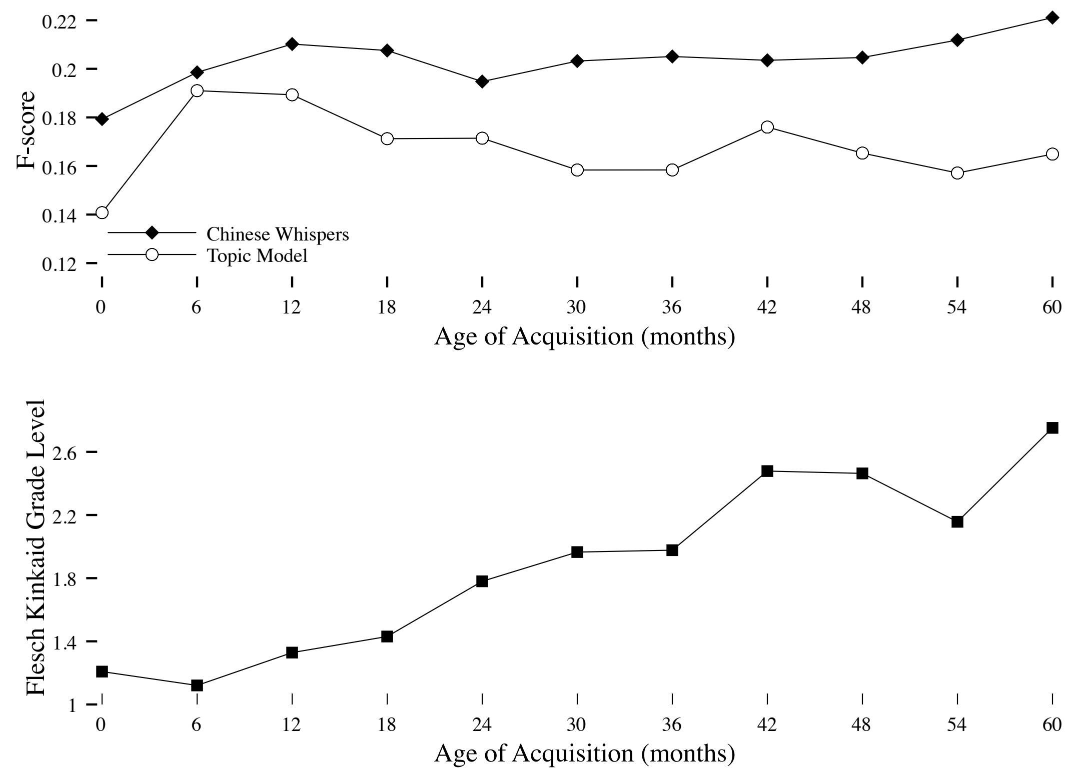


Figure 3.5: Model performance and reading level within CHILDES.

categories, the spoken nature of CHILDES prevented us from making use of such representations (as the parser produced consistently erroneous output). Furthermore, including lexicalised dependencies would be unrealistic, as complex syntactic information is almost certainly not available to children. Parameters for the topic model were set as in the previous experiment. The resulting clusters were evaluated against the gold-standard clusters from Section 3.3.1. Additionally, a complexity measure for each document was computed using the Flesch-Kincaid Grade Level (FKGL) index.[5] FKGL yields a readability score that corresponds to a United States grade level (lower scores mean that the text is easier to read). For example, a score of 8.2 would indicate that the text is expected to be understandable by an average student of age 13–14. One would expect the readability of child-directed speech to be generally low but (crucially) to increase with age.

[5] The FKGL index estimates readability as a combination of the average number of syllables per word and the average number of words per sentence: $\text{FKGL} = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 1.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$.

3.3.2.2 Results

Figure 3.5 shows how model performance (F-score) varies with age together with the average reading level (FKGL) of speech directed at each age group. For clarity the results in Figure 3.5 are grouped into six-month bins. As can be seen the performance of the two models mirrors that of the previous experiment (Section 3.3.1). CW yields better F-scores compared to the topic model; its performance also improves with more data. We next examined the linear relationship between the FKGL readability index and the models' output as measured by F-score using correlation analysis. CW and FKGL were significantly correlated (Pearson's $r = 0.572$, $p < 0.05$). The topic model did not correlate significantly with the FKGL or with Chinese Whispers. These results corroborate the findings of Section 3.3.1 – CW outperforms the topic model in terms of F-score *and* seems to more faithful simulate infant category learning.

Inspecting the output of the topic model's output suggests that its poor performance stems from an inability to produce an appropriate number of categories. With low γ settings the model exhibited a very strong preference for grouping words into no more than one or two topics; at higher settings it exhibited the reverse preference, organising words into extremely fine-grained categories. We attribute this effect to the extreme amount of noise present in the CHILDES data, and the lack of coherence within documents. The above results thus use the same parameter settings as in preceding experiments.

3.3.3 Experiment 4: Incremental Category Acquisition

While the previous experiments explored how effectively the two models capture large-scale category information it did not assess the effect of incrementality. The difficulty in performing such an evaluation is that it requires a snapshot of category structure throughout the process of category acquisition. Getting such snapshots from children would be ideal, however a longitudinal study of category acquisition would be a major undertaking spanning several years. Getting such snapshots from adults is also problematic, as they clearly possess a great deal of world knowledge about the target words used in a hypothetical experiment. As a compromise we conducted a study in which participants were given a series of paragraphs containing nonsense words and then asked, after having read each paragraph, to group the nonsense words into categories. Our hope was that the use of nonsense words would prevent adults from employing any previously-acquired word knowledge to which they might have access, making their process for constructing new categories more similar to that of a child's. If so, such a study would illuminate the kinds of interim categories the mind might construct when presented with minimal information about a set of novel stimuli.

3.3.3.1 Method

Thirteen source documents were compiled from Wikipedia articles on various technical domains, including medicine, physics, biology, and mixology[6]. Each document consisted of 3–5 paragraphs, each containing between 4–6 sentences in which a small number of re-occurring content words were replaced with nonce words (nine on average per document). Figure 3.7 shows an example document presented to participants; a full list of documents annotated with nonce words is provided in Appendix B. The study was completed by 250 participants, mostly undergraduates from the University of Edinburgh.

One serious concern in conducting a study like this is ensuring that participants do not actually perform a separate, but related, task in which they instead determine the mapping between nonsense words and their meaningful equivalents. We mitigated this problem by extracting the text from highly technical documents, the subject matter of which would almost certainly be unfamiliar to participants and thus limiting the amount of world knowledge they could bring to bear. Also of concern was avoiding priming subjects with the number of categories; to avoid such influence, participants were asked to group target words into clusters by dragging items together on a virtual canvas, rather than by assigning labels or placing items into pre-specified bins. A snapshot of the experimental interface our participants saw is given in Figure 3.6.

The topic model and CW were trained on the same set of paragraphs, and the interim clustering produced after processing each document saved, in order to investigate how well the models captured the interim categories formed during incremental learning. Note that both models were trained from a blank state, reflecting a lack of pre-existing world knowledge. Again, we used a bag-of-words representation for CW as the prevalence of nonsense words in the data resulted in many parsing mistakes. Following on Experiment 1, we then applied the topic model and CW to the same set of paragraphs and evaluated the resulting categories against those produced by participants, again using F-score (Agirre and Soroa 2007).

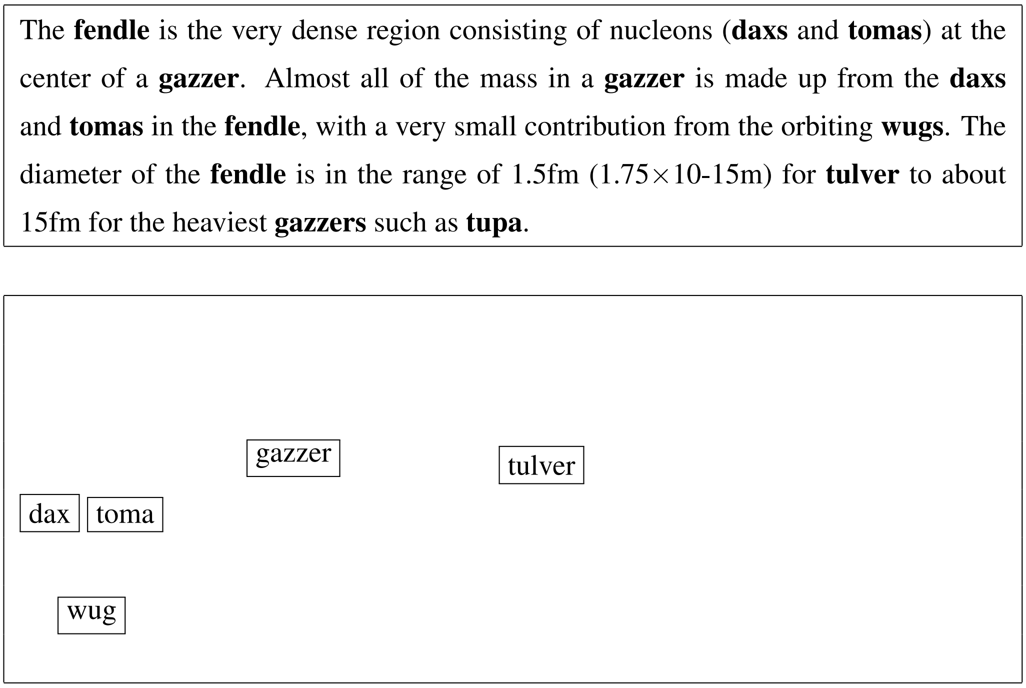
[6] Molecular Mixology is the term applied to the process of creating cocktails using the scientific equipment and techniques of molecular gastronomy.

In this study you'll be given a series of 4-8 short paragraphs to read. The paragraphs contain several highlighted nonsense words; your job is to figure out which of these words belong to the same category. You may decide that all the words belong to the same category (they're all the same kind of thing), that each belongs to its own category (they're all different things), or anything in between (some are one kind, some are another).

After you read each paragraph, use your mouse to group the words below it into whichever categories you think are appropriate. Once you have grouped the words into categories, click ”next” to get the next paragraph.

This first paragraph is just an example; after reading it you might decide that it talks about two kinds of things, fruits and vegetables, and decide to group the words ”apple”, ”orange”, and ”pear” into one category, and the words ”lettuce” and ”spinach” into the other.

As you work your way through the experiment, words you've already seen will turn from blue to grey. You can still move these words around, though, so if you decide that your previous grouping was incorrect, you can (and should!) change it. Don't worry if you can't understand what the paragraphs are talking about. The words will be scrambled; do your best to figure out which (nonsense) words belong together anyway!



The **fendle** is the very dense region consisting of nucleons (**daxs** and **tomas**) at the center of a **gazzer**. Almost all of the mass in a **gazzer** is made up from the **daxs** and **tomas** in the **fendle**, with a very small contribution from the orbiting **wugs**. The diameter of the **fendle** is in the range of 1.5fm (1.75×10-15m) for **tulver** to about 15fm for the heaviest **gazzers** such as **tupa**.

Figure 3.6: The incremental categorisation task as seen by participants in Experiment 4. Each trial consisted of a series of paragraphs from the same source document; the words to be clustered (shown in boldface) remained constant, with participants asked to update their clustering after each trial.

1 In physics, the word annihilation is used to denote the process that occurs when a subatomic **fendle** collides with its respective anti**fendle**. Since **blicket** and **tulver** must be conserved, the **fendle**s are not actually made into nothing, but rather into new **fendle**s. Anti**fendle**s have exactly opposite additive quantum numbers from **fendle**s, so the sums of all quantum numbers of the original pair are zero.

2 Hence, any set of **fendle**s may be produced whose total quantum numbers are also zero as long as conservation of **blicket** and conservation of **tulver** are obeyed. When a low-**blicket** **dax** annihilates a low-**blicket** **tupa** (anti**dax**), they can only produce two or more gamma ray **toma**s, since the **dax** and **tupa** do not carry enough mass-**blicket** to produce heavier **fendle**s and conservation of **blicket** and linear **tulver** forbid the creation of only one **toma**. These are sent out in opposite directions to conserve **tulver**.

3 However, if one or both **fendle**s carry a larger amount of kinetic **blicket**, various other **fendle** pairs can be produced. The annihilation (or decay) of an **dax** **tupa** pair into a single **toma** cannot occur in free space because **tulver** would not be conserved in this process. The reverse reaction is also impossible for this reason, except in the presence of another **fendle** that can carry away the excess **tulver**.

Figure 3.7: A sequence of paragraphs presented to participants in Experiment 4. Paragraphs were drawn from `http://en.wikipedia.org/wiki/Annihilation`, with select content words replaced with nonsense words.

3.3.3.2 Results

Firstly, we assessed how well our participants agreed on the category acquisition task.[7] We computed the F-score of a single participant's clustering for each phase as the average F-score between it and each of the other participants' clusterings for that phase; and then calculated the mean reliability as the average F-score of all phases for all participants. On the category acquisition experiment, participants achieved a mean reliability of 0.694. Interestingly, inter-annotator agreement for the final of five-paragraph documents is significantly lower, likely due to differences in task difficulty between source documents. CW achieved a comparable F-score of 0.656, followed by the topic model with an F-score of 0.634. These F-scores were computed by a procedure similar to the human reliability described above. The model was treated as a single participant and the F-score for each stage was computed as the average F-score between the model's clustering in that stage and each participant's clustering, with the individual stage scores averaged to produce the final score.

Figure 3.8 shows the F-scores achieved by the two models for each phase against the human upper bound. It is interesting to note that both models are close to human performance, with Chinese Whispers having mostly the lead over the topic model. Interestingly, inter-annotator agreement drops precipitiously at the final stage; based on an inspection of participants' clusters after 4 and 5 paragraphs it appears that many participants extensively revised their clusters after encountering the final paragraph. Feedback from participants corroborates this observation, with a number of participants commenting on their tendency to 'second-guess' themselves after encountering the final paragraph.

3.4 Discussion

At first glance the scores on the large-scale task (Section 3.3.1) for both models appear quite low. Our aim in this first experiment, however, was merely to establish a comparison between the two approaches on a clustering task. This is challenging considering that the models are expected to assign 500+ words into an unspecified number of well-defined semantic categories from word co-occurrence information alone. Humans acquire semantic categories from a richer environment based on their sensorimotor experiences in addition to linguistic input.

[7] Subject data for the experiments described in Section 3.3.3 is available from `http://bit.ly/categorization`.

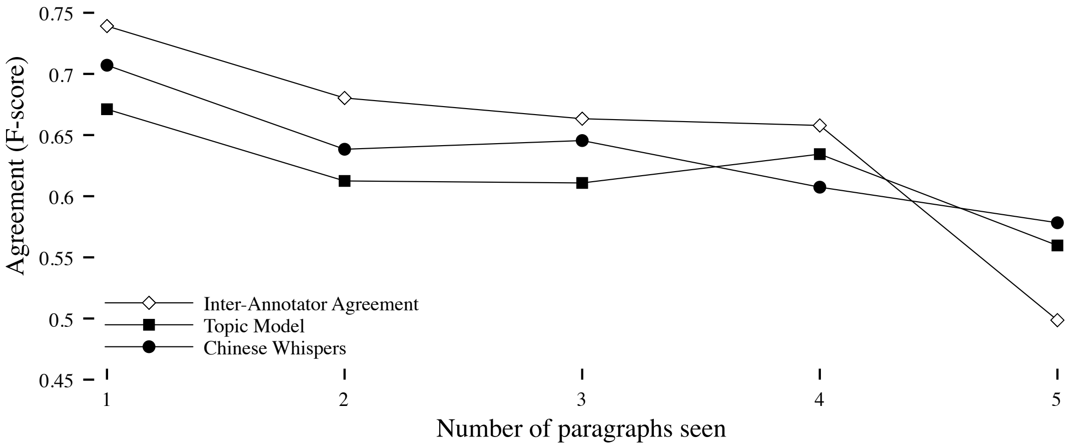


Figure 3.8: Model performance and inter-annotator agreement after encountering N paragraphs, averaged across documents. Note that F-score is here used to compute the similarity of performance between participants rather than to assess the accuracy of individual models or participants. As a result, we should expect F-score to decrease over time as participants' categorisations diverge from one another in the face of increased information.

Regardless, a strict comparison of results shows that CW outperformed the topic model on this large-scale category experiment. Manual inspection of the clusters output by the topic model suggests an explanation: the learnt topics, while clearly capturing some notion of semantic relatedness between words, rarely correspond to the desired semantic categories. Instead they cut across categories, collating words that share a theme or context rather than words belonging to a common category. The clusters output by CW, conversely, capture more of the semantic category information but tend to do so at a higher level (e.g. conflating FRUIT, VEGETABLE, and FOOD into a single meta-category).

This is particularly interesting in light of the differences between the two models; CW is a simpler model, both in terms of the way it represents and forms categories. Recall that the algorithm creates a unipartite graph with one type of nodes (i.e. words) which can be interconnected freely. In the topic model, semantic information is organized in a bipartite graph consisting of words, topics, and their interconnections. This more structured representation does not seem appropriate for the category acquisition task. In particular, the notion of topic as it is used in the context of the topic model is not equivalent to that of a semantic category. The relative success of CW, combined with its simplicity and plausibility, suggests that such comparatively simple models can often provide a better approach for modeling low-level cognitive tasks. From this observation we reach two conclusions: for one, it is likely a mistake to conflate the notion of a 'topic' as it is used in the context of a topic model with that of a semantic category.

The results from Section 3.3.2 involving child-directed speech lend themselves to a similar conclusion; CW learns categories which are closer to the gold standard and presumably more closely aligned with those learnt by children during acquisition. While we obviously have no access to this actual category structure, it is reasonable to assume that it grows in complexity proportionally to the complexity and quantity of the input to which a child is exposed. Additionally, there are numerous parallels that could be drawn between such a graph-based model of category representation and the results of research into category-specific deficits in patients with cognitive impairments (Tyler et al. 2000). While not something we have explored here, a graph-based representation could be used to simulate category-specific defects or localised memory loss.

The results of the third experiment show that CW (and the topic model to a lesser extent) produce categories *incrementally* that are both meaningful and cognitively plausible. Interestingly, in this experiment the upper bound (i.e. inter-annotator agreement) is high despite the seeming difficulty of the task[8]. This suggests that people are quite consistent in the types of categories they form even when those categories are based on only one or two pieces of information, and enforces the idea that, in the absence of real-world knowledge, people learn categories in an incremental fashion (Lamberts and Shapiro 2002).

Both of the models explored in this chapter model acquire categories as flat, non-overlapping clusters; when new concepts are encountered they are placed into a single category based on their similarity to existing exemplars and their relative position in the semantic network. Neither model considers and sort of higher-order relationship between categories. To introduce such relationships we need to model hierarchical category structure, in which we consider both relations between exemplars and categories (e.g. *apple* and *orange* belonging to the category FRUIT) and relations between pairs of categories (e.g. FRUIT and VEGETABLE belonging to the supercategory FOOD). Inspection of the clusters produced by participants in Section 3.3.3 reveals that, even

[8] While conceptually unsurprising, we nevertheless found this result somewhat unexpected given the number of complaints from participants regarding the difficulty of the task.

when presented with explicit instructions to organise words into clusters, participants often chose to organize words into hierarchies rather than flat categories. While it is far from clear that the cognitive mechanisms for organising concepts into categories are inherently hierarchical (Sloman 1998), it seems equally apparent that flat clusters are insufficient for capturing cognitive performance on categorisation tasks. In the following chapter we explore the tendency of our participants to spontaneously generate hierarchical structures in greater depth and demonstrate a model of hierarchical category acquisition which draws inspiration from the graph-based Chinese Whispers model presented here.

Chapter 4

Incremental Category Acquisition Using Hierarchical Random Graphs

In the previous chapter we presented a model of natural language category acquisition and demonstrated its ability to learn meaningful categories from text. This model was similar to the majority of existing categorisation models in that it learnt a flat category structure, i.e. it organised concepts into clusters but described no explicit relationship between those clusters. Under our approach in the previous chapter concepts were grouped into hard clusters, e.g. in which an exemplar *tomato* belonging to category FRUIT may not also belong to category VEGETABLE. An alternate approach would have been to organise concepts into soft clusters, e.g. in which *tomato* may belong to both FRUIT and VEGETABLE to a varying degree. Neither representation describes the relationships between categories (e.g. FRUIT and VEGETABLE both belong to an overarching category PLANTS). In this chapter we attempt to address this failing by introducing a model of category acquisition which learns an hierarchical structure over a set of concepts. While it acquires significantly more complex representations than its predecessor this new model requires no additional input or supervision (i.e. it operates on the same input as the previous, flat model), making the task considerably more difficult than before.

A major focus of the flat category acquisition model of the previous chapter was on maintaining *incrementality* and *non-parametricity*, two key properties we had identified as being essential for maintaining cognitive plausibility. These constraints are carried forward into this chapter and are addressed using similar techniques to those of Chapter 3. We describe a non-incremental version of our model along with the process for evaluating its output and assess the model's performance on a number of datasets using both corpus- and feature-based representations. We then present a modified version of the model which maintains incrementality and evaluate this variant against incrementally constructed hierarchies produced by human participants in an elicitation study. We conclude by highlighting the difficulty of this task and the strength of the hierarchical model by exploring the disagreement between participants in the elicitation study.

4.1 Related Work

Traditional models of category acquisition tend to fall into one of two classes: exemplar models and prototype models. In an exemplar model categories are represented by a set of previously encountered instances of members of the category (Medin and Schaffer 1978, Nosofsky 1988); novel concepts are categorised through comparison with the stored exemplars for each category. Conversely, in a prototype model categories are represented by a single prototypical instance (Reed 1972, Rosch 1973) and categorisation of novel concepts is performed through comparison with only the prototype for each category. While the greater flexibility of exemplar models has been often shown to provide a significant advantage (Nosofsky 1998), a great deal of recent work has focused on categorisation models in between these two extremes. The Adaptive model of Anderson (1990) and the Varying Abstraction Model of Vanpaemel and Storms (2008) both represent categories using clusters of multiple exemplars; other work has represented categories using multiple prototypes (Verbeemen et al. 2007).

All of these models, however, address only the horizontal aspect of categories (Rosch 1978) – they are of limited use when applied to the task of determining category membership at differing levels of abstraction. Varying levels of abstraction in category definitions, along with relationships between categories (e.g. the hyponym-hypernym relationships between DOG, MAMMAL, and ANIMAL), have seen less study. Griffiths et al. (2007a) apply an hierarchical Dirichlet process to the task of category acquisition, but stop short of applying their two-level solution to the task of inferring hierarchical structure, instead using a nested hierarchiy of Dirichlet processes to interpolate between exemplar and prototype representations. Somewhat more applicably, Palmeri (1999) and Verheyen et al. (2008) apply traditional exemplar based models of category acquisition to the task of inferring relationships between categories. Their results indicate that participants' preference for so-called basic-level categories (Rosch 1978), previously believed to be more-or-less constant, varies strongly across categorisation tasks and with respect to the order in which concepts are encountered.

From a more computational standpoint a great deal of research has focused on the task of inferring hierarchical structure from text, often in the form of inducing task-specific lexical taxonomies. Semantic resources like WordNet, a lexical taxonomy organised according to various semantic relations (Fellbaum 1998), have proven invaluable across a wide variety of natural language processing tasks. As a consequence there exist a large number of systems addressing aspects of the taxonomy induction task, from term extraction (identifying a list of concepts to be taxonomised; Kozareva et al. (2008)) to term relation discovery (identifying semantic relations, e.g. IS-A, between terms; Hearst (1992); Berland and Charniak (1999)) and fully automatic taxonomy construction (creating an hierarchical structure over a set of terms; Kozareva and Hovy (2010); Navigli et al. (2011)). Unlike the model presented here, these approaches tend to be supervised or semi-supervised, and almost universally operate directly on text corpora rather than some form of intermediate representation. Furthermore, they often discard the structural flexibility of recent categorisation models (e.g. Griffiths et al. (2007a), Vanpaemel and Storms (2008)), preferring to impose a pre-determined structure on the induced taxonomy. They also tend to reflect a worldview in which there exists a single 'correct' hierarchical organisation for a particular set of terms. While these assumptions are valid in the context in which these systems are traditionally presented (the automatic extension or inference of WordNet-like resources), they render the systems cognitively implausible and limit their application to the task of modelling human category acquisition.

Traditional NLP approaches to the task of taxonomy induction ignore the concept of *incrementality*, the idea that human category learning is not segmented into disjoint 'training' and 'testing' phases but rather takes place alongside concept acquisition and category application. An incremental learner possessing a set of concepts has *at all times* a category structure that explains those concepts to a greater or lesser extent. As she encounters new concepts they are added to this expanding structure; as the category structure is used to make predictions about concepts (either novel or previously observed) it is refined and updated. After encountering a great many concepts and making many category-inspired predictions, the incremental learner will likely possess a similar category structure to that of a non-incremental ('batch') learner. Crucially, however, she will have been able to make use of this structure from the observation of her very first concept. Behavioural evidence (Bornstein and Mash 2010) suggests that the incremental learner more accurately models the process by which human learners acquire new concepts and leverage category structure to predict properties of those concepts.

In the following section we describe a model of hierarchical category acquisition which respects these constraints: our approach is unsupervised, operates on a graph-based representation, and uses a model averaging technique to avoid imposing structural bias on the acquired hierarchy. While the output of the model is a single organisation of concepts into a hierarchical structure, this final hierarchy represents a consensus drawn from a forest of multiple, equally-valid organisations. In our initial experiments we employ a non-incremental version of the model, but it is trivially adapted to the task of incremental learning and we explore the effect of such an adaptation in subsequent experiments.

4.2 Hierarchical Random Graphs

The Hierarchical Random Graph (HRG) model has been applied to the construction of hierarchical representations for a wide variety of structured and unstructured network data: a bacterial metabolic network, a food-web among grassland species, and a network of associations amongst terrorist cells (Clauset et al. 2008). In a more light-hearted application, Clauset et al. (2007) illustrated the appeal of the HRG's implicit model selection technique by reconstructing the high-level conference structure between (American) college football teams from a network of matches. Despite the demonstrable flexibility of the model and its clear successes on a wide variety of data, the only language-related application of which we are aware concerns word sense induction. Klapaftis and Manandhar (2010) create a graph of contexts for a polysemous target word and use an HRG model to organise them hierarchically under the assumption that differences in tree height between contexts correspond to differing levels of sense granularity. The advantages of the HRG model in our context mirror theirs; traditional category acquisition models, like their counterparts in word sense disambiguation, generally impose a flat structure on their target data. Additionally, the HRG operates on an intermediate, graph-based representation (rather than directly on corpora) and requires no outside supervision.

Inference using Clauset et al.'s HRG model can be described as a two-step process. In the first step the model repeatedly samples dendrograms from the space of possible binary trees over a set of concepts using probabilities derived from a graph describing the pairwise relationships between those concepts. Once this sampling process has reached convergence (i.e. subsequent samples no longer result in any signficant improvement in fit), the model samples an ensemble of structures, each representing a plausibly-fitting dendrogram, and merges this ensemble to produce a *consensus hierarchy*. While each structure in the post-convergence ensemble (and, indeed, the ‘best-fit’ dendrogram sampled at convergence) describes a binary tree the application of this final step allows the model to determine the most appropriate hierarchical structure without imposing a significant bias towards particular structures (i.e. preferring flat clusters or deep, n-ary trees.). This consensus process can also be viewed as a form of model selection (Holyoak 2008) in which the HRG uses input data to not only infer a suitable structure but also the form that structure should take.

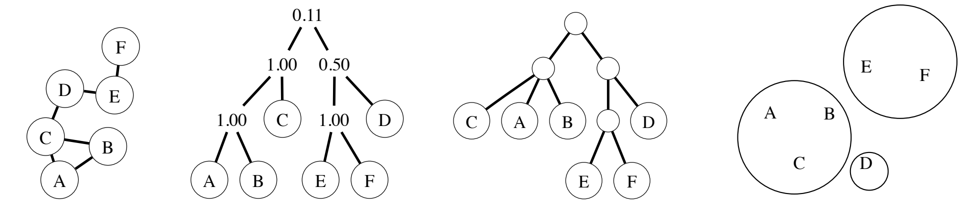
This approach differs from hierarchical clustering in that it explicitly acknowledges that most real-world networks have many plausible hierarchical structures and does not therefore seek a single hierarchical representation for a given network. It also fits well with the nature of category acquisition, in which there rarely exists a uniquely correct categorisation that fully explains a set of concepts. Rather, it is most often the case that different categorisations of the same concepts are appropriate for different tasks and criteria.

More formally, an HRG consists of a binary tree and a set of likelihood parameters, and operates on input organised into a *semantic network*, an undirected graph in which nodes represent terms and edges between nodes indicate a relationship between pairs of terms (Figure 4.1a). From this representation, the model constructs a binary tree (a *dendrogram*) whose leaves correspond to nodes in the semantic network (Figure 4.1b); the model then employs a simple Markov chain Monte Carlo (MCMC) process in order to explore the space of possible binary trees and derives a consensus hierarchical structure from the ensemble of sampled models (Figure 4.1c).

4.2.1 Representing Hierarchical Structure

Formally, consider a semantic network $S = (V,E)$, where $V = \{v_1, v_2 \ldots v_n\}$ is the set of vertices, one per term, and E is the set of undirected edges between terms in which $E_{a,b}$ indicates the presence of an edge between v_a and v_b.

Given a network S the HRG constructs a binary tree D whose n leaves correspond to V and whose $n-1$ internal nodes denote an hierarchy over V. Because the leaves remain constant for a given S, define D as the set of internal nodes $D = \{D_1, D_2 \ldots D_n\}$ and associate each edge $E_{a,b} \in E$ with an internal node D_i being the lowest common



(a) Input graph (b) Binary tree (c) Hierarchy (d) Clusters

Figure 4.1: Flow of information through the Hierarchical Random Graph algorithm. From a semantic network (4.1a), the model constructs a binary tree (4.1b). Edges in the semantic network are then used to compute the θ parameters for internal nodes in the tree; the maximum-likelihood-estimated θ parameter for an internal node indicates the density of edges between its children. This tree is then resampled using the θ parameters (4.1b) until the MCMC process converges, at which point it can be collapsed into a n-ary hierarchy (4.1c). The same collapsing process can be also used to identify a flat clustering (4.1d).

parent of $a,b \in V$. The core assumption underlying the HRG model is that edges in S have a non-uniform and independent probability of existing. Each possible edge $E_{a,b} \in E$ exists with a probability θ_i, where θ_i is associated with the corresponding internal node D_i.

For a given internal node D_i, let L_i and R_i be the number of leaves in D_i's left and right subtrees, respectively; let E_i be the number of edges in E associated with D_i (colloquially, the number of edges in S between leaves in D_i's left and right subtrees). For each $D_i \in D$, we can estimate the maximum likelihood for the corresponding θ_i as $\theta_i = \frac{E_i}{L_i R_i}$. The likelihood $\mathcal{L}(D,\theta|S)$ of an HRG over a given semantic network S is then given by:

$$\mathcal{L}(D,\theta|S) = \prod_{i=1}^{n-1} (\theta_i)^{E_i}(1-\theta_i)^{L_i R_i - E_i} \tag{4.1}$$

4.2.2 Sampling Hierarchical Structures

Because the space of possible dendrograms that can be constructed over a set of n concepts is super exponential with respect to n (Clauset et al. 2007), the HRG makes use of a Markov chain Monte Carlo (MCMC) process (Algorithm 3) to determine the most appropriate dendrogram for a given semantic network. During each iteration of this process the algorithm randomly selects a node within the current dendrogram;

Algorithm 3: MCMC Sampling for Hierarchical Random Graphs

```
Compute the likelihood $\mathcal{L}(D,\theta)$ of the current binary tree.
Pick a random internal node $D_i \in D$.
Randomly permute $D_i$ according to Figure 4.2, yielding a modified dendrogram $\hat{D}$.
Compute the likelihood $\mathcal{L}(\hat{D},\theta)$ of this new dendrogram.
if $\mathcal{L}(\hat{D},\theta) > \mathcal{L}(D,\theta)$ then
    accept the transition;
else
    accept with probability $\mathcal{L}(\hat{D},\theta)/\mathcal{L}(D,\theta)$.
end
Repeat;
```

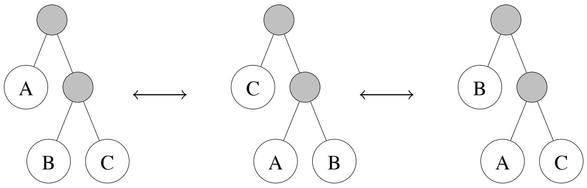


Figure 4.2: Subtree permutations used by the HRG's sampling process. Any internal node with subtrees A, B, and C can be permuted to one of two possible alternate configurations. Shaded nodes represent internal nodes which are unmodified by such permutation.

the subtree rooted at this node is permuted according to Figure 4.2. If the permutation improves the overall likelihood of the dendrogram the transition is accepted and the process repeats. If the permutation fails to improve the overall likelihood of the dendrogram it may be accepted according to standard Metropolis acceptance rules, i.e. with a probability proportional to the difference in likelihood ($\mathcal{L}(\hat{D},\theta)/\mathcal{L}(D,\theta)$) between dendrograms.

4.2.3 Inferring a Consensus Hierarchy

Once the MCMC process has converged the model is left with a dendrogram over the terms from the input semantic network. As in standard hierarchical clustering,

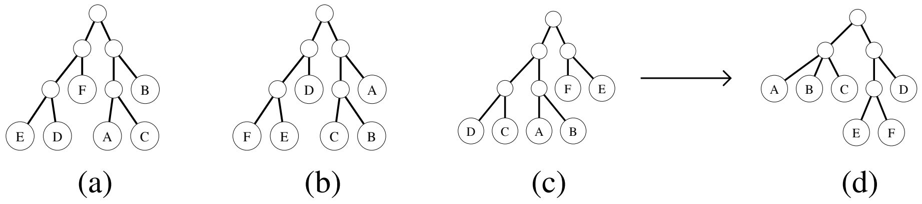


Figure 4.3: An illustration of the consensus hierarchy procedure. From the three sampled dendrograms on the left we identify the clusters encoded in each dendrogram; dendrograms (a),(b) and (c) encode the clusters $\{AC, ABC, DE, DEF, ABCDEF\}$, $\{BC, ABC, EF, DEF, ABCDEF\}$, and $\{AB, CD, ABCD, EF, ABCDEF\}$, respectively. Rejecting those clusters which do not appear in a majority of dendrograms leaves the clusters $\{ABC, EF, DEF, ABCDEF\}$, which are encoded in the consensus hierarchy (d).

however, this imposes an arbitrary structure which may or may not correspond to the observed data — the dendrogram at convergence will be similar to an ideal binary tree given the graph, but a binary tree may not be the most plausible organisation of concepts. Indeed, for an hierarchical categorisation task it is quite unlikely that a binary tree will provide the most appropriate representation of the relationships between concepts.

To avoid encoding such bias we employ a model averaging technique to produce a *consensus hierarchy*. For each of a set of dendrograms sampled after convergence, this process first identifies the set of possible clusters encoded in the dendrogram, e.g. the dendrogram in Figure 4.1b encodes the clusters $\{AB, ABC, EF, D, DEF, ABCDEF\}$. As in Clauset et al. (2007), each cluster instance is then weighted according to the likelihood of the originating HRG (Equation 4.1); it then sums the weights for each distinct cluster across all resampled trees and discard those whose aggregate weight is lower than 50% of the total observed weight. The remaining clusters are then used to reconstruct an hierarchy in which each subtree appears in the majority of trees observed after the sampling process has reached convergence (hence the term consensus hierarchy).

4.2.4 An Incremental Adaptation of the Hierarchical Random Graph Model

Due to the use of a random sampling technique for inference and, more significantly, the reliance on a fully-formed input graph), the HRG as described is *not* an incremental model. In this section we describe a set of modifications to the Hierarchical Random Graph model that allow it to learn an hierarchical structure in an incremental fashion. Because of its statistical nature, i.e. the use of an MCMC process to search the space of structures, these modifications are less straightforward than those required to incrementalise the Chinese Whispers model (see Chapter 3, Section 3.2.1).

Algorithm 4: An incremental version of the Hierarchical Random Graph algorithm.

```
Initialise an empty semantic network $S$ and dendrogram $D, \theta$.
for each instance of a concept $c$ do
    if $c \in S$ then
        Update $S$ with modified similarities;
    else
        Add $c$ to $S$;
        Identify the leaf node $l = \arg\max_{n \in D} sim(c, n)$;
        Create a new internal node $n$ having $c$ and $l$ as children;
        Replace $l$ with $n$ in $D$;
    end
    Resample $D, \theta$ using Algorithm 3.
end
```

Where a batch version of the HRG begins with a fully-formed semantic network derived from an (unspecified) external source and a randomly initialised dendrogram over the concepts therein, an incremental version by definition begins with an empty network and a correspondingly empty dendrogram. As novel concepts are encountered they are added to both the semantic network, with edges connecting to similar, previously encountered concepts, and to the dendrogram. The dendrogram is updated to include the new concept by identifying the existing concept to which it most similar and replacing that concept in the dendrogram with a node possessing both the new and existing concepts as children. As new evidence concerning existing concepts is encountered the semantic network is updated with new similarities (possibly adding or removing edges between concepts) and the dendrogram resampled to reflect these changes. A more formal description of this process is presented in Algorithm 4.

Modified to operate in this fashion the HRG can be accurately described as an incremental learner: it possesses at all times an hierarchical structure over exactly the set of concepts it has encountered, and this structure can be used to make predictions about a concept immediately after encountering it for the first time. Unfortunately, the nature of the HRG means that this structure will necessarily be a binary dendrogram. To overcome this limitation we repeat the consensus hierarchy step of the HRG after each iteration (i.e. after encountering each new piece of evidence).

4.3 Evaluating Inferred Hierarchies

Evaluation of taxonomically organised information is notoriously hard (Hovy 2002) due to the inherently subjective and application-specific nature of the task (e.g. a dolphin can be a mammal to a biologist, but a fish to a fisherman or someone visiting an aquarium). Defining a 'gold-standard' organisation against which we can evaluate an induced hierarchy is consequently difficult; while we can (and do!) make use of manually constructed ontologies such as Wordnet (Fellbaum 1998), we would ideally like to evaluate the output of our model against a wide range of equally valid hierarchies.

Assuming that we have identified a gold-standard hierarchy (or, again, a set of equally valid hierarchies), it remains unclear how we should assess the degree to which an hierarchy produced by our model captures the relational knowledge encoded in the gold standard. Computational approaches to taxonomy induction often employ a task-based evaluation in which a model-induced hierarchy is evaluated based on its accuracy at predicting specific relations, e.g. IS-A, between concepts (Yang and Callan 2009). Alternatively, one could employ more tree-theoretic measures like tree edit distance (Demaine et al. 2009) or tree alignment (Bille 2005) to directly compute the similarity between hierarchies. We take the latter approach and define a novel measure for scoring the similarity between arbitrary hierarchies over an identical set of leaves; this measure is then used to assess how well the output of the model matches a gold-standard hierarchy.

Even assuming that we possess a gold-standard hierarchy and a matching evaluation metric we're still unable to make meaningful comparisons to previous work; most research on category acquisition, especially incremental category acquisition, model the task as one of acquiring flat (rather than hierarchical) category structure (Fountain and Lapata 2011). To enable comparisons against this work we need a means of evaluating our induced hierarchical category structure as a flat clustering. To do this we define an heuristic method for collapsing HRG output into clusters and employ the cluster F-score measure described in Chapter 3 (Cluster F-score; Agirre and Soroa

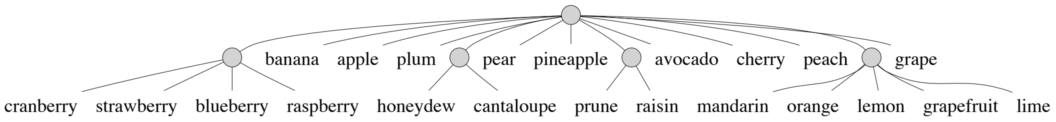


Figure 4.4: A subset of the Wordnet gold-standard hierarchy for the concept *fruit*. While the concept contains a moderate number of exemplars (21), these are organised into a relatively flat hierarchical structure, with sub-concepts corresponding (roughly) to *berry*, *melon*, *dried fruit*, and *citrus*.

(2007)).

4.3.1 Obtaining a Gold-Standard Hierarchy

To assess the hierarchies produced by the HRG model it was necessary to construct a gold-standard hierarchy over the set of concepts provided to the model. We obtained such an hierarchy by extracting the `is-a` relationship tree from WordNet (Fellbaum 1998), taking only concepts and relationships involving words appearing in the McRae et al. norms. Concepts appearing in the norms but absent from WordNet were removed from the set of target words provided to the HRG, and thus not assessed; a total of 493 concepts appear in both.

Specifically, we first identified the full hypernym path in WordNet for each noun in McRae et al.'s 2005 dataset, e.g. *Apple* > *Plant Structure* > *Natural Object* > *Physical Object* > *Entity*. These hypernym paths were then combined to yield a full taxonomy over McRae et al.'s concepts; internal nodes having only a single child were recursively removed to produce a final, compact taxonomy containing 186 semantic classes (e.g. *animals*, *weapons*, *fruits*) organized into varying levels of granularity (e.g. *songbirds* > *birds* >*animals*). Because the McRae et al. concepts are limited to concrete nouns, this final taxonomy is rooted in the WordNet *Physical Object* concept. A visualisation of a subset of this taxonomy appears in Figure 4.4; the full taxonomy is encoded in Appendix C.

4.3.2 Flat Cluster Evaluation

To evaluate a flat clustering into classes we use the F-score measure introduced in the SemEval 2007 task (Agirre and Soroa 2007); like traditional F-Score it is defined as the harmonic mean of precision and recall. In the context of a cluster evaluation, precision

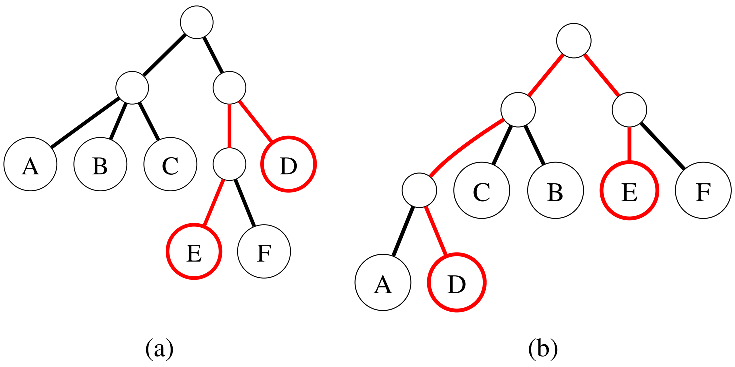


Figure 4.5: An illustration of the tree-height correlation metric. For each pair of leaf nodes, we compute the walk distance between nodes in each hierarchy. The walk distance between D and E is 3 in (a) and 5 in (b); the tree-height correlation between (a) and (b) is 0.518.

is defined as the number of correct members of a cluster divided by the number of items in the cluster; recall is defined as the number of correct members of a cluster divided by the number of items in the gold-standard cluster.

Because the output of the consensus hierarchy procedure is an hierarchy rather than a hard clustering, it is necessary to perform an additional post-processing step (Algorithm 5) in which this hierarchy is flattened into a simple clustering. This can be done in a straightforward, principled fashion using the HRG's θ parameters. For a given $\mathcal{H}(D,\theta)$ this process identifies internal nodes whose θ_k likelihood is greater than the mean θ and who possess no parent node whose θ_k likelihood is also greater than the mean. Each such node is the root of a densely connected subtree; each such subtree is then assumed to represent a single discrete cluster of related items, where $\bar{\theta} = mean(\theta)$ (illustrated in Figure 4.1c).

4.3.3 Hierarchical Evaluation

Although informative, an evaluation based solely on F-score puts the HRG model at a comparative disadvantage as the task of hierarchy induction is significantly more difficult than simple, flat clustering. To overcome this disadvantage we propose an automatic method of evaluating taxonomies directly by first computing the walk distance between pairs of terms that share a gold-standard category label within a gold-standard and a candidate hierarchy, and then computing the pairwise correlation between distances in each tree (Lapointe 1995). This correlation captures the intuition that a 'good'

Algorithm 5: Flattening the output of an HRG into a hard clustering

```
Let $D_k$ be the root node of $D$.
if $\theta_k > \bar{\theta}$ then
    output the leaves of the subtree rooted at $D_k$ as a cluster
else
    repeat 2 with left and right children of $D_k$.
end
```

hierarchy is one in which items appearing near one another in the gold hierarchy also appear near one another in the induced one. It is also conceptually similar to the task-based IS-A evaluation (Snow et al. 2006) which has been traditionally used to evaluate hierarchy.

Formally, let $G = \{g_{0,1}, g_{0,2} \cdots g_{n,n-1}\}$, where $g_{a,b}$ indicates the walk distance between terms a and b in the gold standard hierarchy. Similarly, let $C = \{c_{0,1}, c_{0,2} \ldots c_{n,n-1}\}$, where $c_{a,b}$ is the distance between a and b in the candidate hierarchy. The *tree-height correlation* between G and C is then given by Spearman's ρ correlation coefficient between the two sets. All tree-height correlations reported in the following experiments were computed using the WordNet-based gold-standard hierarchy over McRae et al.'s 2005 nouns described in Section 4.3.1.

4.3.4 Comparison Models

To provide a reference for comparison it is necessary to establish a set of baselines against which the HRG can be evaluated. Because to the best of our knowledge there does not exist a similar model for acquiring hierarchical categories from a graph-based intermediate representation, we compare hierarchies inferred by the HRG against those produced by three models from similar tasks.

Because the task of inferring hierarchical categories follows naturally from the previous chapter's task of inferring flat categories, we make a comparison against the Chinese Whispers model described therein. Because the two models infer different structures some manipulation is necessary in order to facilitate a comparison; this took the form of heuristically flattening the HRG's inferred hierarchies into discrete clusters and comparing the results against those inferred by Chinese Whispers. Hierarchies inferred by the HRG model were flattened into discrete clusters using Algorithm 5.

Chinese Whispers and the HRG operate on identical input, a semantic network of

Algorithm 6: Standard bottom-up agglomerative clustering

```
Let $K$ be the set of identified categories
Initialise $K$ to include one category for each $k$ concept
while $|K| > 1$ do
    Let $a, b = \arg\max_{a,b} sim(a, b)$
    Remove $a$ and $b$ from $K$
    Create a tree node $t$ with $a$ and $b$ as children
    Add $t$ to $K$
end
```

concepts and relations; both models operate by positing an initial organisation over the input concepts and iteratively improve this organisation until reaching consensus. This iterative process differs between the two: Chinese Whispers performs an iterative step by performing a randomised colour propagation between nodes in the input graph, whereas the HRG employs a more probabilistic, MCMC approach (i.e. permute the current representation, re-compute the likelihood after permutation, and decide to accept or reject the permutation according to the increase or decrease in likelihood). Additionally, both models are unsupervised in as much as they take no additional input beyond the graph to be categorised. External information can of course be introduced by encoding it in the graph, but both algorithms are completely agnostic as to the *source* of the input. Neither model makes use of oracle category-exemplar pairs, seed clusters, or other techniques for introducing supervision. Given these similarities we would expect the two models to perform similarly on a flat cluster evaluation, i.e. the hierarchical categorisation induced by the HRG should capture roughly the same boundaries between categories as the flat clustering produced by Chinese Whispers.

Our second baseline is Brown et al.'s 1992 agglomerative clustering algorithm that induces a mapping from word types to classes. The Brown et al. algorithm begins with K categories for the K most frequent concepts and proceeds by alternately adding the next most frequent concept to the category set and merging the two categories which result in the least decrease in the mutual information between class bigrams. The result is an hierarchy over categories with clustered concepts at the leaves. Because the Brown et al. algorithm relies on direct corpus counts and takes as input a set of documents rather than a semantic network, we also compare against a standard agglomerative clustering technique (Sokal and Michener 1958, Algorithm 6). This technique produces a binary dendrogram in a bottom-up fashion by recursively identifying concepts or categories (subtrees) with the highest pairwise similarity.

4.4 Large-Scale Experiments

In the following experiments we evaluate the HRG on a series of hierarchy induction tasks. Because HRGs provide a means of inducing an hierarchy over graph-based input representations and is not directly affected by the manner in which these graphs are produced, these experiments were designed to investigate how differences in the topology and quality of the input graph influence the algorithm's performance.

Experiments 5 and 6 investigate the quality of the hierarchies induced by the HRG when provided high- and low-quality input, respectively. Experiment 7 then investigates a novel approach to increasing the quality of the input network without introducing external supervision. Following these three experiments we conduct an assessment of how well the HRG reconstructs a forest of human-produced hierarchies (Experiment 8).

All four of these experiments evaluate the HRG and comparison models on a non-incremental task – given a gold-standard hierarchy (or a set of gold-standard hierarchies, as in Experiment 8), how well does the model recreate that hierarchy from a semantic network? Given our stated interest in developing *incremental* models of category acquisition, we thus conduct an experiment evaluating the HRG against hierarchies produced by human participants during an incremental categorisation task (Experiment 9). This experiment is similar in structure and procedure to the analogously incremental experiment in the preceding chapter, Experiment 4.

4.4.1 Experiment 5: Inducing Hierarchies From Featural Representations

Because the HRG takes as input an abstract, graph-based representation we first consider the case in which the input graph provides high-quality information about the similarity of concepts. Inducing such a graph automatically, e.g. using corpus-based similarities or a relation-extraction algorithm, would likely introduce errors which could complicate an assessment of the HRG. Because the HRG performs only the hierarchy induction step of a (hypothetical) complete pipeline for concept extraction and categorisation, we conduct our initial experiment to assess its performance when

Method	F-score	Tree Correlation
HRG	**0.507**	**0.168**
CW	0.464	—
Agglomerative	0.352	0.137

Feature Norms

Table 4.1: Cluster F-score and tree-height correlation for the HRG and baseline models using as input a semantic network constructed over McRae et al.'s 2005 nouns and feature-based similarities.

provided with known good input.

To conduct such an assessment we constructed a semantic network using similarities derived from the feature norming study of McRae et al. (2005). Each noun was represented as a vector with dimensions corresponding to the possible features generated by participants in the norming study; the value of a term along a dimension was taken to be the frequency with which participants generated the corresponding feature when given the term. For each pair of terms an edge was added to the semantic network if the cosine similarity between their vector representations exceeded a fixed threshold (set to 0.15 and tuned empirically on held-out concepts).

The resulting network was then provided as input to the HRG and the resulting dendrogram resampled until it reached convergence. The binary tree at convergence was again resampled to produce both a consensus hierarchy and a set of flat clusters (according to the procedure described in Section 4.3.2). The resulting consensus hierarchy was evaluated by computing the tree-height correlation between it and the gold-standard (WordNet-derived) hierarchy; the resulting clusters were evaluated by computing the cluster F-score using a gold-standard (human-produced) clustering (See Chapter 2, Section 2.4).

Unfortunately, the Brown et al. algorithm operates on corpora rather than a semantic network (or any similar representation that could be derived thereof). As a result it was omitted from the results listed in Table 4.1. Additionally, we only report cluster F-score for the Chinese Whispers model as it does not induce an hierarchical clustering. When evaluated using F-score, the HRG algorithm produces better quality clusters compared to Chinese Whispers, in addition to being able to organise them hierarchically. It also outperforms agglomerative clustering by a large margin; a similar pattern emerges when the HRG and agglomerative clustering are evaluated on tree correlation.

The taxonomies produced by the HRG are a better fit against the WordNet-based gold standard; the difference in performance is statistically significant ($p < 0.01$) using a t-test (Cohen and Cohen. 1983).

Manual inspection of the induced hierarchies suggests an explanation for the increased performance of the HRG relative to Chinese Whispers on the flat clustering evaluation, a task on which we might logically expect the latter model to outperform the former. Because Chinese Whispers is restricted to a flat clustering and cannot extract fine-grained relations within clusters, it produces a flat categorisation at a conceptually higher level (e.g. identifying a cluster corresponding to *animal* rather than separate clusters for *mammal*, *fish*, and *bird*). The more flexible nature of an hierarchical categorisation, combined with the heuristic method used to collapse it into a flat clustering, allows the HRG to make more fine-grained distinctions when possible. Of course, the gain from a fine-grained hierarchical organisation of concepts is limited; the agglomerative clustering, which by definition must produce a binary tree, encodes fine-grained distinctions within logical clusters where none may exist. Forcing the model to identify for each concept a single concept to which it is most similar produces irrelevant or incorrect distinctions for categories possessing very flat internal structure.

4.4.2 Experiment 6: Inducing Hierarchies from Corpus-based Representations

As the semantic network derived from the McRae et al. feature norms provides what could plausibly be described as oracle similarities the results of Experiment 5 can be considered as an upper bound of sorts for what can be achieved by the HRG when provided with perfect or near-perfect input. Feature norms capture detailed knowledge about word meaning and concept relationships which would be difficult if not impossible to obtain from corpora. Given this, it is interesting to explore how well the HRG performs on an hierarchy induction task when provided with a lower-quality semantic network.

To conduct this assessment we extracted concept similarities using co-occurrence statistics computed from the British National Corpus (Burnard and Aston 1998) and provided the resulting semantic network to the HRG, Chinese Whispers, and agglomerative models (as in Experiment 5); additionally, we employed the algorithm of Brown et al. (1992) to induce an hierarchy directly from the corpus. Unfortunately, this algo-

Method	F-score	Tree Correlation
HRG	**0.276**	0.104
CW	0.274	—
Brown	0.258	**0.124**
Agglomerative	0.122	0.077

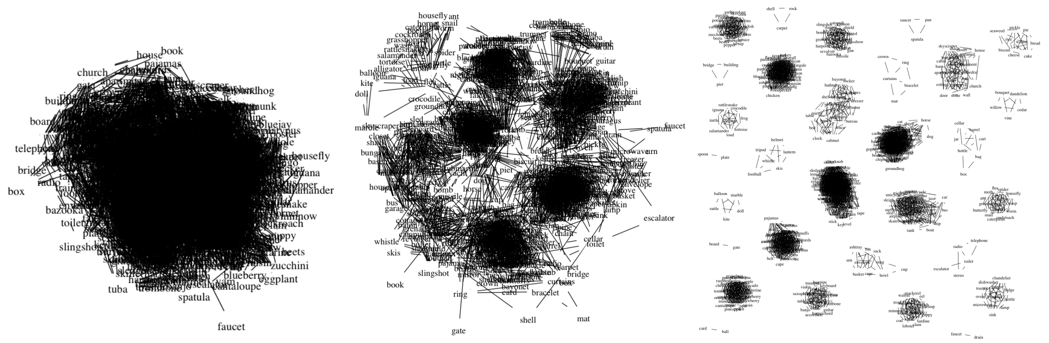
Corpus Similarities

Table 4.2: Cluster F-score and tree-height correlation evaluation for hierarchies inferred over McRae et al.'s 2005 nouns; all algorithms are run on the BNC. Note that tree correlation for the Chinese Whispers (CW) model is not reported, as it produces only a flat clustering and not an hierarchy.

rithm requires the number of desired output clusters to be specified in advance; in all trials this parameter was set to the number of clusters in the gold-standard clustering (41), thus providing the Brown-induced hierarchies with a non-trivial oracle advantage.

Again as in Experiment 5, target words were represented as vectors in a semantic space, but with dimensions corresponding to possible co-occurring context words; the concepts from McRae et al. (2005) were again used as the set of target words. To construct vector representations we extracted context windows of five words on either side of each occurrence of a target word along with 5,000 vector components corresponding to the most frequent non-stopwords in the BNC. Raw frequency counts were transformed using pointwise mutual information to produce the final representations. These representations were then used to construct a semantic network in which each node corresponded to a target word; an edge was added between a pair of target words if the cosine distance between their vectors exceeded a pre-defined threshold. To facilitate a comparison with Experiment 5 this threshold was set to 0.15.

This semantic network was then input to all three models (HRG, Chinese Whispers, and agglomerative clustering). We evaluated the resulting hierarchies, along with the Brown-induced hierarchy produced directly from the corpus, against the same gold-standard flat and hierarchical clusterings used in Experiment 5. Results for all four categorisations are shown in Table 4.2. On the flat clustering evaluated (listed in the table as F-score) the HRG has a slight advantage against both Chinese Whispers and the Brown et al. algorithm; differences in their performance are not, however, statistically significant. Standard agglomerative clustering is the word-performing method, with a decrease in F-score of approximately 1.5. On the tree-height correlation evaluation the



(a) $s = 0.0$ (b) $s = 0.5$ (c) $s = 1.0$

Figure 4.6: A semantic network as derived from the BNC (a) and the same network reweighted using a flat clustering produced by CW (b). As s approaches 1.0 the network exhibits an increasingly strong small-world property, eventually reconstructing the input clustering only (c).

HRG is comparable to Brown; both algorithms are significantly better ($p < 0.01$) than agglomerative clustering.

4.4.3 Experiment 7: Inducing Hierarchies from Small-World Graphs

Performance of the HRG is better when the semantic network is based on feature norms (compare Tables 4.1 and 4.2), both in terms of tree-height correlation and F-score. This suggests that the algorithm is highly dependent on the quality of the semantic network used as input. HRGs are known to operate well on so-called ‘small-world’ networks – graphs composed of densely connected subregions with relatively few edges between them (Kleinberg 2000, Clauset et al. 2007). While the feature-based semantic network input to the HRG in Section 4.4.1 could be accurately described as a small-world graph, inspection of the corpus-induced network from the previous section (see Figure 4.6a) shows it to be emphatically *not* a small-world graph. To determine what effect (if any) the small-world structure has on the hierarchy ultimately induced by an HRG we repeated the experiment of Section 4.4.2 using a set of corpus-induced semantic networks with an artificially-imposed small-world structure.

To impose such structure we first obtained a set of (flat) clusterings over the concepts in the network. These clusterings were taken from the output of the Chinese Whispers (see Chapter 3 or Biemann (2006) for details) and Brown et al. algorithms (The agglomerative clustering baseline was omitted due to both its purely hierarchical nature and its abysmal performance; see Table 4.2). Notably, neither of these algo-

Method	F-score	Tree Correlation
HRG	0.276	0.104
HRG + CW	**0.291**	0.161
HRG + Brown	0.255	**0.173**

Reweighted Corpus Similarities

Table 4.3: Cluster F-score and tree-height correlation evaluation for taxonomies inferred by the HRG using semantic network derived from the BNC and re-weighted using CW and Brown.

rithms requires any oracle information (e.g. requiring the specification of the correct number of categories to infer) and thus do not introduce any outside supervision into the overall hierarchy induction task.

The process used to impose a small-world structure onto the existing semantic network is relatively straightforward. We compute a modified weight $\widehat{W}_{A,B}$ between a pair of terms A,B according to Equation (4.2), where s indicates the proportion of edge weight drawn from the clustering, $W_{A,B}$ is the edge weight in the original (BNC) semantic network, and $C_{A,B}$ is a binary value indicating that A and B belong to the same cluster (i.e. $C_{A,B} = 1$ if A and B share a cluster; $C_{A,B} = 0$ otherwise).

$$\widehat{W}_{A,B} = (1-s)W_{A,B} + sC_{A,B} \tag{4.2}$$

The value of the s parameter was tuned empirically on held-out development data and set to $s = 0.4$ for both CW and Brown algorithms. Each re-weighted network was then used as input to the HRG model, with the resulting taxonomies evaluated in the same manner as in Section 4.4.2.

Table 4.3 shows results for cluster F-score and tree-height correlation for the HRG when using a graph derived from the BNC without any modifications (i.e. the semantic network inferred from the BNC and used in Section 4.4.2), as well as two re-weighted versions using the CW and Brown clustering algorithms, respectively. As can be seen, re-weighting improves tree-height correlation substantially: HRG with CW and Brown is significantly better than HRG on its own ($p < 0.05$). In the case of CW, cluster F-score also yields a slight improvement. Interestingly, the tree-height correlations obtained with CW and Brown are comparable to those attained by the HRG when using the human-produced feature norms (differences in correlations are not statistically

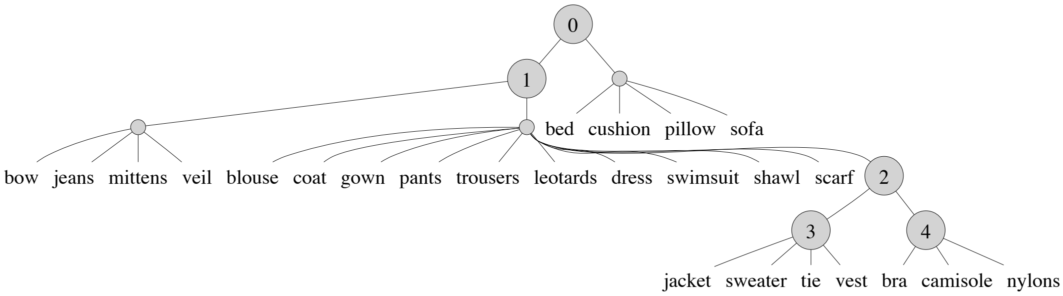

0
1
2
3
4
bed cushion pillow sofa
bow jeans mittens veil blouse coat gown pants trousers leotards dress swimsuit shawl scarf
jacket sweater tie vest bra camisole nylons

Figure 4.7: An excerpt from an hierarchy induced by the HRG using the BNC semantic network with Brown re-weighting. The HRG does not provide category labels for internal nodes of the hierarchy, but subtrees within this excerpt correspond roughly to (0) *Textiles*, (1) *Clothing*, (2) *Gendered Clothing*, (3) *Men's Clothing*, and (4) *Women's Clothing*.

significant). An excerpt of an HRG-induced hierarchy is shown in Figure 4.7.

The increase in quality of the resulting hierarchy after re-weighting is particularly exciting given that the resources used to both estimate the original graph and perform the re-weighting can be automatically obtained from corpora. Constructing high-quality representations of word meaning using feature norms is costly and time consuming; these results show that we can approximate word meaning using only distributional semantics while retaining comparable performance.

4.5 Estimating Human Performance

In the previous experiments we evaluated the HRG against a gold-standard hierarchy derived from WordNet. While this provides a decent assessment of how well the model recreates a particular hierarchical structure it fails to address two of our primary motivations for using the HRG in the first place: that a model of category acquisition should operate incrementally and should not impose a bias in favour of a particular taxonomic structure. By evaluating against a WordNet-derived hierarchy we implicitly assume that the correct (or rather, *a* correct) hierarchy over our set of target words should have the same or similar structure (in terms of tree depth or granularity) to WordNet; by performing a only single evaluation based on a fully-specified semantic network we ignore any impact of incremental learning. Evaluating only the final induced hierarchy also leaves open the question of how well the model reflects the *process* of category acquisition rather than merely the end result.

To address these oversights we conducted a pair of experiments, one to gauge performance against a set of non-expert, human-produced hierarchies and one to gauge performance on an incremental category acquisition task. Both experiments took the form of a web-based elicitation study to obtain human-produced hierarchicies followed by an evaluation of the HRG (and baselines where possible) on the same task. Section 4.5.1 describes the study and evaluation against human-produced hierarchies; Section 4.5.2 makes use of the same framework to obtain and evaluate a series of incrementally produced hierarchies.

4.5.1 Experiment 8: Computing a Human Upper Bound

Experiments 5 and 6, in which the HRG and baselines were assessed against a WordNet-derived hierarchy, demonstrate the strengths of the HRG model for batch category acquisition but fail to address the concern that for a given set of concepts there exist multiple, equally correct hierarchies describing the relationships between concepts. This concern stems from the observation that multiple plausible hierarchies may describe an entirely correct view of the relationships between concepts and categories given differing criteria for assessing category membership; even the organisation of biological species into a single taxonomy, perhaps the canonical example of an hierarchical organisation of concepts into categories, is under constant revision (Cavalier-Smith 2004). For the concepts used in the previous experiments the WordNet hierarchy represents merely one of many equally valid hierarchies. Noting this, it is interesting to explore how well the output of the HRG fits within the set of plausible, valid hierarchies over a fixed set of concepts.

To explore this we conducted an elicitation study in which human participants were presented with a 12-word subset of concepts (*cedar*, *lemon*, *pear*, *tomato*, *peach*, *pigeon*, *owl*, *chicken*, *lion*, *tiger*, *cat*, *dog*, *bear*, *python*) and asked to organise them into arbitrary hierarchies. We then applied the HRG and baselines to induce hierarchies over these same concepts, evaluating each in turn against those produced by participants in the study. The list of concepts was chosen heuristically; we first selected a sub-hierarchy of the WordNet tree (LIVING THINGS) along with its subtrees (e.g. ANIMALS, PLANTS), and chose target concepts from within these trees in order to produce a taxonomy in which some items were differentiated at a high level (e.g. *python* vs. *dog*) and others at a fine-grained level (e.g. *lion* vs *tiger*).

The elicitation study was conducted using Amazon Mechanical Turk[1], and in-

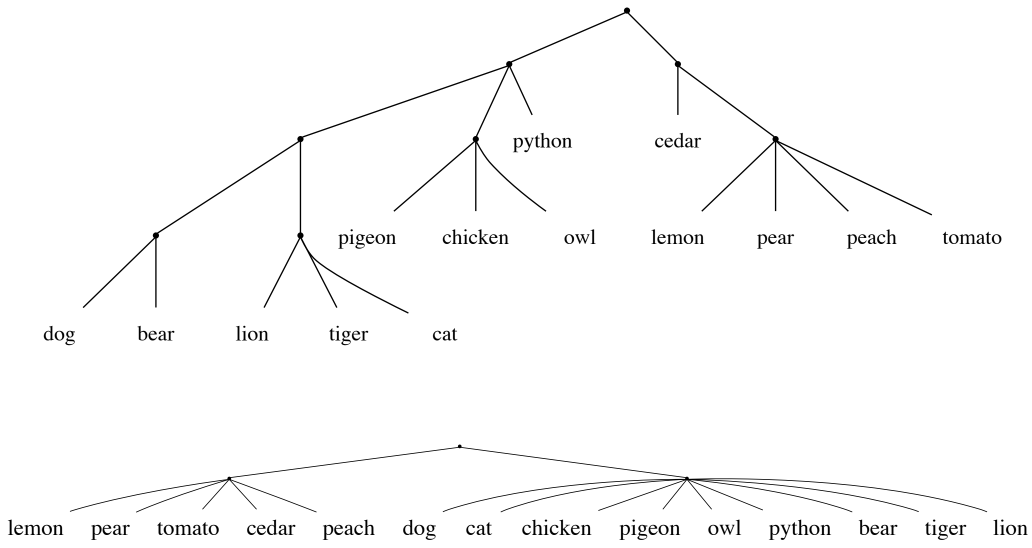


Figure 4.8: Hierarchies produced by two human participants in the 12-word hierarchy elicitation study. Note that both hierarchies describe a ‘correct’ organisation of concepts into categories, with the top hierarchy describing categories at a relatively fine-grained level and the bottom hierarchy capturing only the distinction between PLANTS and ANIMALS.

volved 41 participants from English-speaking countries. No specific guidelines as to what features participants were to use when organising these concepts were provided. Participants were presented with a web-based, graphical, mouse-driven interface for constructing an hierarchy over the chosen set of concepts.

To evaluate the HRG along with the baselines from Section 4.3.4 against the resulting hierarchies we constructed a semantic network over the subset of concepts using similarities derived from the BNC; this network was a subgraph of that used in Section 4.4.2. we also repeated the re-weighting procedure described in Section 4.4.3 using the clusterings induced by the CW and Brown algorithms over the subgraph to update edge weights.

Table 4.4 shows the performance of the HRG and baselines from the preceding sections evaluated against hierarchies produced by participants in the elicitation study. For the models and baselines the score reported is the mean of the tree-height corre-

[1] Amazon’s Mechanical Turk (`http://mturk.com`) is an online ‘marketplace for work’ in which anonymous, non-expert workers complete simple, atomic tasks in exchange for financial compensation. In recent years it has been used to explore a wide variety of cognitive science and text processing tasks (Mason and Suri 2011), including a limited form of supervised hierarchy induction (Dakka and Ipeirotis 2008).

Method	Tree Correlation
HRG	**0.412**
CW	0.324
Brown	0.181
Agglomerative	0.274
HRG + CW	0.307
HRG + Brown	0.155
Inter-annotator Agreement	0.511

Table 4.4: Model performance and inter-annotator agreement on a subset of the target words used in Sections 4.4.1-4.4.3, applied to a subset of the semantic network used in Section 4.4.2. Instead of a WordNet-derived hierarchy, models were evaluated against hierarchies manually produced by participants in an online study. For models the reported score is the mean tree-height correlation between the hierarchy output by the model and those created by participants; inter-annotator agreement is reported as the mean pairwise tree-height correlation between hierarchies created by participants.

lations between the hierarchy output by the model and the hierarchy output by each participant; inter-annotator agreement is the mean pairwise tree-height correlation between hierarchies output by participants. Participants achieve a mean pairwise tree correlation of 0.511, indicating that there is a fair amount of agreement between participants regarding the hierarchical organisation between the specified concepts. The HRG comes close achieving a mean tree correlation of 0.412, followed by Chinese Whispers and agglomerative clustering. For the sake of completeness we also applied the Brown model to the paragraphs read by participants, though the amount of data available was too limited for it to produce meaningful results. In general, the HRG manages to produce hierarchies that resemble those generated by humans to a larger extent than any of the competing algorithms applied. The results in Table 4.4 also hint at the fact that the hierarchy induction task is relatively hard, even for human annotators leveraging real-world knowledge, as participants do not achieve anything like perfect agreement despite the fact that they are asked to taxonomise only 12 words.

As an example, consider the two human-produced hierarchies shown in Figure 4.8 – both describe plausible, 'correct' organisations of the 12 concepts into hierarchical categories, but at wildly different levels of granularity. Both hierarchies capture the intuitive division of concepts into PLANTS and ANIMALS, but the topmost hierarchy encodes significant additional information, identifying sub-categories corresponding to the concepts of FELINES, BIRDS, and FRUITS. Because both capture the same high-level division, however, agreement between these hierarchies is relatively high, at 0.790.

4.5.2 Experiment 9: Evaluating Incrementality

The preceeding three experiments assess the HRG in a standard 'batch-processing' context. All three experiments follow the same general form: construct a semantic network over a fixed set of concepts (e.g. using feature norms or co-occurrence counts), use the HRG to infer an hierarchy over those same concepts, and evaluate the result against some form of gold-standard hierarchical categorisation. While this model provides an effective means of evaluating the HRG's performance on an hierarchical category acquisition task – and enables a valuable comparison against baseline models which operate in the same fashion – it provides no assessment of how well the HRG reflects human performance on this task, as it fails to account for incrementality.

To rectify this we present a novel experiment for assessing the modified model's performance against human category learners in an incremental context. As with the incrementalising modifications to the HRG, this experiment mirrors that of Chapter 3, Section 3.3.3, with significant changes reflecting the difference in acquiring hierarchical (rather than flat) category structure.

4.5.2.1 Capturing Incremental Hierarchy Snapshots

Evaluating an incremental model of category acquisition is a significantly more difficult task than that of evaluating its non-incremental counterpart due to the necessity of obtaining *snapshots* of the gold-standard category structure at intermediate stages of the category acquisition process. Ideally, of course, these snapshots would be obtained from children, so as to reflect the stages of category learning actually experienced by humans; conducting such a longitudinal study of category acquisition would, however, be a serious undertaking. Collecting such snapshots from adults presents other difficulties, as they clearly possess a considerable amount of real-world knowledge encoded in a well-formed and mostly fixed hierarchy.

To obtain these snapshots we conducted an elicitation study using Amazon Mechanical Turk in which partipants were presented with a series of text passages and asked to construct an hierarchy over the concepts therein. In an attempt to overcome the problem of participants' use of real-world knowledge the passages were drawn from technical documents describing complex concepts (e.g. particle physics). Additionally, the target concepts participants were asked to categorise were replaced by nonsense words. This obfuscation should prevent participants from leveraging any domain-specific knowledge they may have.

The passages were drawn from six Wikipedia articles concerning particle physics[2], colour theory[3], and biology[4]. Each document consisted of 3-5 paragraphs, with each paragraph containing between 4-6 sentences in which a small number of re-occurring content words were replaced with nonsense words; an example document with paragraph breaks is shown in Table 4.5 (Appendix D contains the full list of documents and nonsense words). Paragraphs from a randomly selected document were presented in order to each participant in a visual, mouse-driven interface in which they could select and move concepts, group concepts together, and organise groups or concepts into arbitrary hierarchies. The interface was derived from that described in Section 4.5.1; as in that study we found that participants had very little difficulty in using the interface to construct a meaningful hierarchy.

After reading the initial paragraph participants were presented with a list of (nonsense) concepts discussed therein and asked to group them into an hierarchy. After indicating that they had completed the hierarchy to the best of their ability they were then presented with the subsequent paragraph and their just-constructed hierarchy; if the paragraph introduced new concepts these were made available for addition to the hierarchy. At this stage, and after each following paragraph, participants were asked to read the paragraph and update their hierarchy if necessary.

To ensure high quality final hierarchies we required all participants to complete a verification task before allowing them to complete the study proper. In this task participants were presented with a single paragraph[5]and a short list of commonly understood concepts (i.e. without the technical/nonsense obfuscation previously described). They were then asked to construct an hierarchy over these concepts based on the paragraph. Only participants whose hierarchies reflected the obvious divide between concepts (i.e. the distinction between FRUITS and VEGETABLES) were allowed to complete the remainder of the study. While the use of a screening task (or 'qualification test' in Mechanical Turk parlance) may impose some bias by selecting for a particular type

[2] en.wikipedia.org/wiki/Hadron, en.wikipedia.org/wiki/Annihilation, en.wikipedia.org/wiki/Atomic_nucleus, and en.wikipedia.org/wiki/Quark

[3] en.wikipedia.org/wiki/Colorfulness

[4] en.wikipedia.org/wiki/Nucleic_acid

1 A **borograve** is an elementary **mim** and a fundamental constituent of matter. **borograve**s combine to form composite **mim**s called **fendle**s, the most stable of which are **vorp**s and neutrons, the components of atomic nuclei. All **fendle**s except **vorp**s are unstable and undergo **mim** decay.

2 Due to a phenomenon known as color confinement, **borograve**s are never found in isolation; they can only be found within **fendle**s. For this reason, much of what is known about **borograve**s has been drawn from observations of the **fendle**s themselves.

3 There are six types of **borograve**s, known as flavors: **dax**, **blicket**, **tupa**, **zav**, **wug**, and **toma**.

4 **dax** and **blicket** types of **borograve**s have the lowest masses of all **borograve** types. The heavier **borograve**s rapidly change into **dax**s and **blicket**s through a process of **mim** decay. Because of this, **dax**s and **blicket**s are generally stable and the most common in the universe, whereas **tupa**, **zav**, **wug**, and **toma** can only be produced in high energy collisions.

5 A **borograve** of one flavor can transform into a **borograve** of another flavor only through the weak interaction, one of the four fundamental interactions in **mim** physics. By absorbing or emitting a boson, any **dax**, **tupa**, or **wug** can change into a **blicket**, **zav**, or **toma**, and vice versa.

Table 4.5: An example document presented to participants in the incremental hierarchy induction task. Numbers on the left indicate the breakdown of the document into paragraphs shown to participants in successive order. This document was originally drawn from the English-language Wikipedia article on quarks, with selected content words (e.g. quark, particle, proton) replaced by nonsense words (borograve, mim, vorp).

of participant it is generally considered a necessary step in conducting crowdsourced studies (Snow et al. 2008, Ipeirotis 2010).

4.5.2.2 Evaluating the HRG Against Incremental Learners

We applied our HRG model to the task of inducing hierarchies for each document presented to participants; the parameter-free nature of the HRG means that it was presented with no more information than that provided to participants in the elicitation study. For each document the HRG was applied to successive paragraphs in the same order as presented to human participants; after encountering each paragraph the semantic network was updated to reflect changes in similarity and to add any newly encountered concepts. The incremental version of the HRG was then resampled until convergence using the updated network. 50 dendrograms sampled post-convergence were then used to construct the consensus hierarchy after each paragraph, resulting in model-induced hierarchies for each document corresponding to those produced by human participants in the elicitation study (i.e. after reading one paragraph, two paragraphs, etc., up to the entire document).

These inferred hierarchies were then compared to the ensemble of hierarchies produced by participants at the corresponding stage; we report the model performance after *N* documents as the mean tree-height correlation between model-induced hierarchies after *N* paragraphs and those produced by participants after reading the same number of paragraphs. For comparison, we also report inter-annotator agreement at each stage as the mean tree-height correlation between participants.

Overall, we find that the HRG model performs relatively well in comparison to human participants on the incremental hierarchy induction task. Its early confusion, in which it tends to initially group all concepts into a single flat cluster, is overcome after a small number of paragraphs. Humans, however, are able to produce comparatively high-quality (agreeing) hierarchies based on very little data, most likely due to their ability leverage existing world knowledge despite our attempts to obfuscate the task. After encountering four or more paragraphs, however, the hierarchies output by the model do not differ significantly from those produced by human participants. As for inter-annotator agreement, we notice that it (perhaps unsurprisingly) differed wildly

[5] "**Apple**, **orange**, and **pear** trees are by far the most popular variety of fruit tree. The fruit they produce, be it gala **apple**s, navel **orange**s, or European **pear**s, is usually sweet, and is preferred for making fresh squeezed juice. Leaf vegetables most often come from short-lived herbaceous plants such as **lettuce** and **spinach**. In many countries, **lettuce** is typically eaten cold and raw in salads, sandwiches, or other dishes. **Spinach** is generally served as a stand-alone side dish or mixed into a salad."

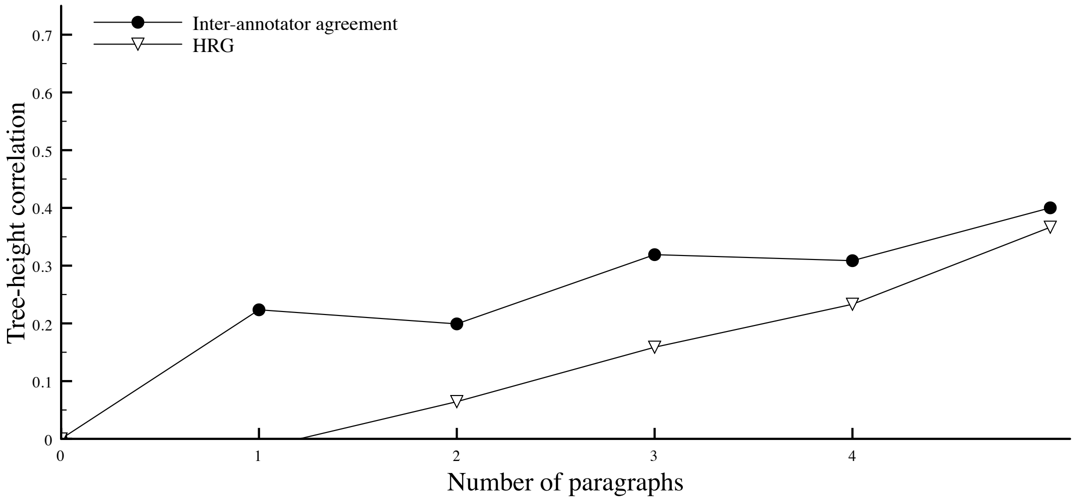
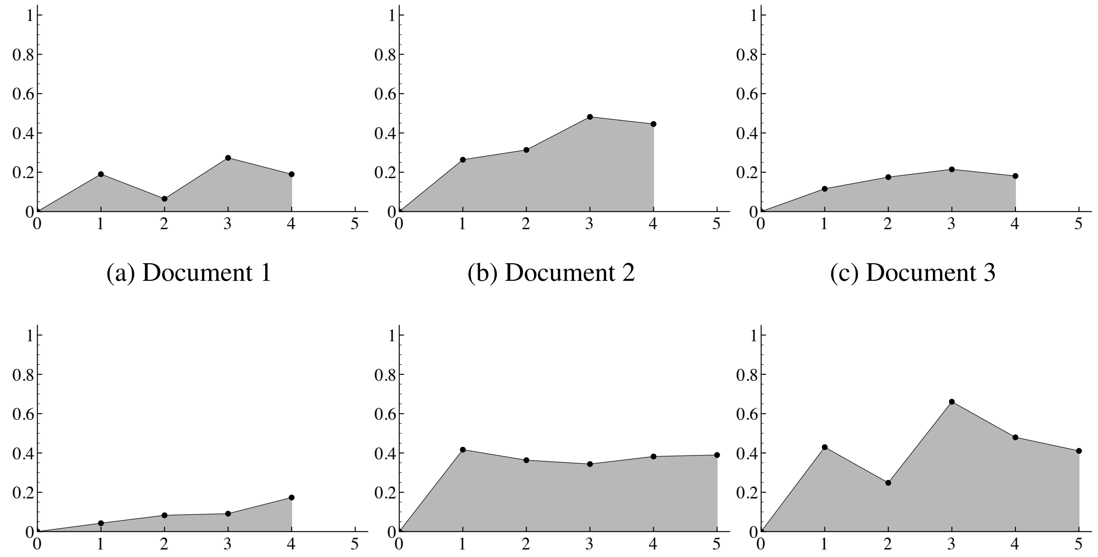


Figure 4.9: Inter-annotator agreement and model performance averaged across documents in an incremental hierarchy induction task. The x-axis represents the number of paragraphs encountered; the y-axis represents the mean pairwise tree-height correlation. For inter-annotator agreement, this corresponds to the mean pairwise correlation between participants; for the HRG it corresponds to the mean correlation between the hierarchy output by the model and that output by each participant after the indicated number of paragraphs.



(a) Document 1

(b) Document 2

(c) Document 3

(d) Document 4

(e) Document 5

(f) Document 6

Figure 4.10: Inter-annotator agreement on across documents in an incremental hierarchy induction task. The x-axis in each plot represents the number of paragraphs encountered; the y-axis represents inter-annotator agreement, computed as the mean tree-height correlation between participants.

across documents. Figure 4.9 reports the inter-annotator agreement averaged across documents; Figure 4.10 reports the inter-annotator agreement for each document. We attribute this discrepancy in agreement to differences in the document; anecdotally, participants often commented on the difficulty of the task, especially when presented with those documents on which would later result in lower inter-annotator agreement.

4.6 Discussion

In this chapter we've presented a model of category acquisition that obeys the constraints set out in Chapter 3, but which induces an hierarchical categorisation rather than a hard, flat clustering. Our hierarchical model is based on Clauset et al.'s (2007) Hierarchical Random Graph algorithm, and operates on an intermediate representation (a semantic network) in an unsupervised fashion. Like the cluster-based algorithm of the preceding chapter, it mirrors the incremental nature human category acquisition and does not require any oracle knowledge about the category structure to be induced (e.g. the number of categories or the structure of the relations between them).

We demonstrated the effectiveness of this model on both high-quality input derived from feature norms and low-quality input extracted from corpus co-occurrence counts, and highlighted the flexibility of the intermediate representation in a re-weighting experiment. Additionally, we illustrated the difficulty of the hierarchical category induction task through a pair of elicitation studies and showed that our HRG-based model provides a good fit for human performance on both incremental and non-incremental hierarchy induction tasks.

Chapter 5

Conclusion

In the preceeding chapters we developed the task of natural language categorisation, exploring corpus-based representations of semantic meaning (Chapter 2 and constructing cognitively plausible models of category acquisition (Chapters 3 and 4). In this final chapter we summarise the primary contributions of this thesis and main findings of each chapter in light of the central claims laid out in the beginning of this thesis. We conclude by discussing a few possible avenues for future work.

5.1 Primary Contributions

The overriding purpose of this thesis has been to push the boundardies of understanding regarding how semantic categories are acquired and applied. Throughout the thesis we have focused on the task of acquiring what we term *natural language categories*, i.e. categories which group words into meaningful semantic clusters; the models and behavioural studies presented herein demonstrate that this category structure is acquired in an incremental, unsupervised, non-parametric fashion by human learners and that this acquisition process can be easily and extensibly modeled as operations on a semantic network. We have shown that incremental, graph-based models can acquire category structure based on distributional semantics and predict the interim categories formed by humans during category learning. Furthermore, we have presented considerable evidence attesting to the incremental nature of category learning in the form of our ‘category snapshot’ studies (Experiments 4 and 9), the design of which is to the best of our knowledge entirely novel.

We hope that a reader who remembers nothing else of this thesis will recall at least the following: that category acquisition is an incremental task that requires no delineated training phase or oracle data, and that models of acquisition must take these properties into account to maintain even the pretense of cognitive plausibility.

5.2 Main Findings

In this thesis we have investigated the extraction and use of corpus-based semantic representations in models of categorisation. Our method of constructing these representations enabled us to explore a series of cognitively plausible models of category acquisition, and to evaluate those models in a novel task designed to highlight their incremental nature. The use of a simple, graph-based intermediate representation in our models enables us to apply them to arbitrary semantic representations, including both traditional feature norms and corpus-derived alternatives. Our models accurately predict category structures constructed by participants during interim stages of an online category acquisition task; the success of their predictions provide further evidence of the incremental nature of human categorisation.

In the following sections we describe this evidence in greater detail, and summarise the main findings of this thesis with respect to the claims put forth in Chapter 1: that corpus-based concept representations can be used in place of traditional feature norms, that models using these representations can be used to predict participants' category structures in an acquisition task, and that the use of an interim, graph-based representation for concepts and categories can greatly facilitate the construction of such models.

5.2.1 Concept Representations

We explored a number of methods for representing the semantics of natural language concepts, all of which provide an unsupervised, automatic alternative to traditional feature norms. Each of these methods define the meaning of a concept based on the contexts in which it frequently appears, but differ in the level at which they define this context (e.g. the words appearing in the same document, topic, sentence, or syntactic relation). To compare methods we conducted a set of experiments in which we provided representations for a large set of natural language concepts to a pair of simple exemplar and prototype models and evaluated the models' performance on common categorisation tasks: category naming, exemplar generation, and typicality rating.

The results of these experiments demonstrate that distributional representations can provide a plausible replacement for traditional feature norms. Unlike feature norms, which are expensive and time-consuming to create, our representations can be automatically extracted, in an unsupervised fashion, from a suitably large corpus. In order to facilitate a comparison against feature-based representations, in our experiments we restricted ourselves to only those concepts for which norming data was available, though there is in theory no obstacle to extracting a list of target concepts simultaneously with their representations.

5.2.2 Category Acquisition

We identified a pair of criteria as essential for defining a cognitively plausible model of category acquisition, *incrementality* and *non-parametricity*. We developed two models meeting these criteria, and evaluated their performance on large-scale and incremental category acquisition tasks. In the latter of these tasks we applied the models to a novel task involving predicting participants' intermediate category structures at various stages during a contrived acquisition task.

Our models were able to construct intermediate category structures that matched those produced by participants, for both flat (Chapter 3) and hierarchical (Chapter 4) categories. Inter-annotator agreement suggests that participants were able to successfully perform the task, regardless of the type of structure elicited, with relatively little difficulty; a comparison of model- and human-induced categories suggests that adult acquisition of unfamilar natural language concepts can be accurately modelled as an incremental, non-parametric task.

5.2.3 Categories as Semantic Networks

We constructed two models of category acquisition in which concepts and categories are represented as nodes and regions (respectively) in a semantic network. The use of this structure to encode category membership greatly facilitated the design of models which met our incrementality and non-parametricity constraings, as well as providing a flexible means of introducing external information into the models' internal representations. We demonstrated this flexibility by using an (automatically-extracted) flat categorisation to influence the construction of an hierarchical categorisation over the same set of concepts, and showed that doing so significantly increased the quality of the resulting hierarchies. We compared the categories output by our models against those learnt by existing hierarchical and flat categorisation models, and showed that they consistently produced higher-quality categorisations of both flat and hierarchical structure.

5.3 Future Work

5.3.1 Hybrid Featural Representations

It is well established that there is considerable overlap in the information encoded in perceptual and linguistic features of concepts (Riordan and Jones 2011) – indeed, this overlap is largely what enables models which rely solely on distributional semantics, like those discussed in this thesis, to construct meaningful semantic representations without access to perceptual stimuli. The inclusion of perceptual features in a vector-based semantic models can improve categorisation performance and allow for the prediction of unobserved features (Johns and Jones 2011).

Future work in this direction involves exploring the impact of including perceptual features, e.g. sense-limited features selected from a norming study or visual features extracted from images, on hierarchical category structure. An interesting question is how perceptual information might be represented in an incremental context, in which certain dimensions may be unobservable, e.g. a child present with visual stimuli (photographs) without accompanying auditory or touch information.

5.3.2 Knowledge-based Categories

In this thesis we have primarily concerned ourselves with exemplar models of categorisation. While we briefly investigated prototype models in Chapter 2, other theories have gone largely unexplored. Classical categories, defined by a rigid list of necessary and sufficient features, are problematic to construct using the tools of natural language categorisation – our use of distributional semantics to represent concept meaning lacks the stringent precision required to define such feature lists, and (other deficiencies aside) it is difficult to imagine classically-defined categories possessing a meaningful hierarchical structure outside of a few narrowly-specified areas (e.g. the biological taxonomy).

It is logical to expect that knowledge- or theory-based categories, however, follow an hierarchical structure and are acquired in an incremental, non-parametric fashion. These categories by definition have a much more complex structure, the acquisition of which entails a great deal of specialised world knowledge. While outside the scope of this thesis, it is not implausible that the techniques presented here could be paired with more complex semantics (e.g. vector-space representations in which dimensions correspond to lexicalised semantic predicates) to produce a model of hierarchical, knowledge-based categories.

5.3.3 Hierarchical Categories

One of our motivations in Chapter 4 for modelling categories using an hierarchical, rather than flat, structure was the inherent difficulty in establishing a single 'correct' categorisation for a given set of concepts. What set of exemplars constitute a particular category changes depending on context and experience (Frassinelli and Lenci 2012); in a culinary context, for example, *tomato* would be considered a VEGETABLE, but in a botanical context it is quite clearly a FRUIT. Furthermore, the level at which people 'cut' an hierarchy of categories into a set of flat categories (what Rosch (1978) calls the 'basic level') differs based on a person's expert knowledge and the task to which category knowledge is applied.

One possible avenue for future work involves the exploration of the different 'correct' structures that can be constructed for a given set of concepts. Roughly speaking, the set of pre-consensus hierarchies sampled by the Hierarchical Random Graph model might represent a starting point for such an exploration. The final consensus hierarchy output by an HRG represents a single categorisation with an arbitrary hierarchical structure; the individual hierarchies sampled to produce this consensus, however, can each be considered an equally plausible categorisation, albeit one with an artificially enforced binary structure. Future work in this direction could investigate the relationship between these interim structures and task-specific or expert categorisations.

The WordNet-based evaluation used in Chapter 4 suggests an obvious additional route for future work. The automatic induction or extension of taxonomic information is a common task in natural language processing, necessitated by the widespread use of and difficulty in creating such taxonomies. Such taxonomies have historically been created using lexical or syntactic patterns, either manually encoded (Hearst 1992, Caraballo 1999) or induced (Snow et al. 2006). The parallels between the task of identifying an hierarchical categorisation for a set of natural language concepts and explicitly inducing a lexical taxonomy (or extending an existing taxonomy such as WordNet) are numerous and obvious; in Chapter 4 we treat a subset of WordNet as a gold-standard categorisation. Adapting the HRG model to produce a more complete WordNet-style taxonomy could provide additional insight into how such taxonomies might be created.

Appendix A

Gold-Standard Category Names and Exemplars

The concepts used in McRae et al. (2005) feature norming study, augmented with category labels produced by participants in an online category naming study described in Chapter 2. The category name under which each concept appears is the most commonly produced label, after accounting for synonymity and differences in spelling.

WEAPON
machete, gun, cannon, harpoon, bayonet, rifle, shield, rocket, bullet, baton, bomb, axe, spear, grenade, slingshot, crossbow, hatchet, wand, pistol, brick, catapult, whip, tomahawk, bazooka, revolver, sword, knife, tank, dagger, shotgun, crowbar, bow, missile

BUG
wasp, moth, worm, grasshopper, beetle, caterpillar, flea, ant, hornet, salamander, housefly, spider, cockroach

COOKWARE
skillet, bowl, toaster, dish, colander, pan, kettle, spatula, saucer, pot

RODENT
hamster, mink, rat, beaver, mole, mouse, gopher, squirrel, otter, chipmunk, bat

CLOTHES

veil, nightgown, sweater, trousers, shoes, dress, gown, jeans, shirt, blouse, camisole, vest, pants

HARDWARE
chain, faucet, drain, hook, door, bathtub, gate, level, doorknob, clamp, drill, toilet, board, cabinet, shovel, pipe, peg, plug, bolts, sink, key, screws, fence

ACCESSORIES
belt, sack, mat, necklace, veil, curtains, crown, napkin, buckle, bookcase, mittens, razor, earmuffs, umbrella, tie, cushion, cap, gloves, fan, cape, bag, bracelet, socks, nylons, scarf, vest, ring, tack, bow, pin

HOUSE
partridge, shack, door, wall, doorknob, basement, cabinet, cellar, cottage, closet

FOOD
coconut, tomato, cucumber, deer, parsley, banana, owl, pickle, walnut, pie, pineapple, eggplant, pumpkin, corn, potato, garlic, mushroom, onions, bread, prune, biscuit, asparagus, yam, rhubarb, blueberry, lettuce, avocado, mink, lobster, celery, tuna, beans, olive, mole, carrot, chicken, cabbage, rice, peach, cheese, beets, pepper, peas, cranberry, lemon, radish, snail, trout, shrimp, lime, cake, mandarin, crab, turnip, grapefruit, grape, zucchini, seaweed, cherry, broccoli, clam, raspberry, strawberry, cauliflower, spinach, sardine, honeydew, plum, raisin

HOME
carpet, mat, beehive, napkin, trailer, ashtray, curtains, bucket, door, hose, doorknob, house, drapes, bedroom, cage, urn, board, plug, hut, apartment, chandelier, cellar, fan, shell, key, cabin, bungalow

TOOLS

screwdriver, knife, key, pipe, pencil, cork, scissors, starling, elevator, hatchet, microscope, crane, spade, ladle, whip, drill, tap, comb, thermometer, level, razor, grater, wrench, fan, doorknob, ruler, shed, peg, racquet, armour, bolts, pliers, typewriter, clock, axe, screws, saddle, hose, bucket, stereo, candle, tray, hoe, projector, tomahawk, crayon, wand, tape, pin, paintbrush, clamp, spear, budgie, plug, chain, crowbar, tack, stick, corkscrew, shovel, thimble, pen, skillet, pot, wheelbarrow, hammer, wheel, blender, brush, machete, tongs, colander, rake, broom, sledgehammer, umbrella, hook, strainer, rattle, sandpaper, anchor, rope, chisel

GARMENT
veil, sweater, apron, gown, jacket, swimsuit, robe, shirt, parka, tie, cloak, blouse, cap, cape, vest, scarf, coat

HOUSING
beehive, tent, trailer, shack, candle, basement, hut, apartment, cellar, cottage, bungalow, cabin, inn, barn

UTENSIL
broom, grater, skillet, whistle, comb, strainer, hose, pen, razor, dish, colander, plate, scissors, spoon, pencil, whip, fork, board, thimble, tripod, brush, spatula, cup, ladle, knife, tongs, crayon, thermometer, pin

ENCLOSURE
sack, box, bathtub, bucket, gate, jar, bottle, cage, barrel, bin, bag, sink, shed, fence, closet

FURNITURE
rocker, bed, sofa, bayonet, shelves, clock, desk, bathtub, bookcase, drapes, stool, bureau, dresser, cushion, chair, cabinet, table, couch, cupboard, bench, armour, mirror, fence

INSECT

wasp, moth, worm, snail, grasshopper, beetle, flea, caterpillar, ant, hornet, butterfly, housefly, spider, cockroach

CONTAINER
sack, box, tray, ashtray, bowl, bucket, jar, envelope, urn, barrel, basket, cage, bottle, cap, bin, pan, cup, bag, mug, tank, pot

FISH
sardine, eel, perch, goldfish, cod, whale, octopus, lobster, guppy, shrimp, mackerel, seal, walrus, salmon, catfish, squid, minnow, salamander, tuna, trout, clam

VEHICLE
bus, raft, boat, unicycle, yacht, trailer, bike, train, rocket, truck, sled, submarine, van, jet, subway, helicopter, trolley, skateboard, tractor, wagon, ship, motorcycle, ambulance, canoe, jeep, buggy, limousine, airplane, elevator, sleigh, dunebuggy, taxi, car, surfboard, tricycle, tank, scooter

REPTILE
iguana, rattlesnake, tortoise, leopard, crocodile, turtle, toad, frog, salamander, alligator

FRUIT
mandarin, nectarine, grape, raisin, plum, grapefruit, tangerine, lime, avocado, rhubarb, prune, apple, walnut, olive, yam, banana, peach, orange, cherry, lemon, strawberry, pumpkin, cranberry, honeydew, tomato, pear, coconut, cantaloupe, blueberry, pineapple, raspberry

OBJECT
tray, cart, broom, budgie, tent, napkin, chain, candle, wall, door, plate, card, cork, pyramid, bouquet, toilet, cage, board, pipe, corkscrew, shell, stick, book, rock

TOY

doll, unicycle, rocker, rattle, sled, football, ball, baton, slingshot, wand, card, skateboard, buggy, marble, balloon, kite, sleigh, surfboard, tricycle, crayon

KITCHEN
tray, skillet, grater, strainer, faucet, bowl, bucket, dish, colander, dishwasher, pan, pickle, sink, cupboard, spatula, cup, kettle, ladle, knife, mug, tongs, pot

APPLIANCE
tray, skillet, stove, microwave, faucet, stereo, toaster, plate, radio, clamp, oven, fridge, toilet, lamp, dishwasher, chandelier, corkscrew, sink, telephone, fan, kettle, blender, freezer, mixer, thermometer

CLOTHING
necklace, belt, veil, pajamas, dress, nightgown, sweater, trousers, boots, leotards, shoes, gown, buckle, apron, earmuffs, jacket, swimsuit, hose, mittens, jeans, shirt, robe, slippers, shawl, parka, cloak, tie, blouse, helmet, bra, cap, camisole, skirt, socks, gloves, cape, nylons, vest, scarf, armour, coat, pants, bow

STRUCTURE
tray, garage, beehive, perch, door, basement, pyramid, cage, hut, cellar, cottage, building, bridge, fence, pier

THING
carpet, gun, box, doll, bike, chain, rifle, bullet, baton, sled, desk, umbrella, envelope, catapult, fork, basket, pipe, fan, shell, building, stick, tongs, rock, tricycle, stone, car, mirror, tack, tank, screwdriver, bow

DEVICE

key, radio, cork, shield, elevator, whistle, microscope, whip, tap, boat, box, fan, tripod, doorknob, door, raft, sword, peg, barrel, armour, telephone, typewriter, clock, microwave, saddle, bucket, stereo, drain, freezer, hoe, projector, crossbow, couch, wand, clamp, tent, escalator, closet, cup, plug, pistol, oven, baton, trailer, stove, mug, cage, jar, napkin, skillet, pot, lamp, house, slingshot, cart, wheel, blender, brush, missile, mixer, keyboard, toilet, parka, bomb, toaster, broom, sledgehammer, catapult, cathedral, umbrella, strainer, buckle, cape, rope, chisel

PLANT
cedar, rhubarb, vine, porcupine, parsley, willow, bouquet, mushroom, dandelion, seaweed

BIRD
budgie, pigeon, stork, peacock, seagull, starling, pheasant, crow, woodpecker, bluejay, parakeet, emu, flamingo, hawk, canary, partridge, oriole, perch, robin, chicken, turkey, duck, rooster, buzzard, platypus, owl, dove, sparrow, ostrich, penguin, nightingale, raven, vulture, birch, finch, blackbird, falcon, swan, salmon, goose, eagle, chickadee, pelican, crane

TRANSPORTATION
raft, boat, unicycle, cart, bus, budgie, escalator, yacht, horse, bike, train, truck, rocket, skis, submarine, van, jet, subway, helicopter, trolley, sailboat, ambulance, canoe, ship, motorcycle, jeep, buggy, wagon, donkey, limousine, elevator, airplane, dunebuggy, taxi, sleigh, wheel, car, tank, scooter

ANIMAL

duck, chickadee, cougar, gopher, hornet, hawk, beaver, hose, cockroach, beehive, camel, bull, goldfish, pickle, eel, giraffe, beetle, leopard, sheep, lobster, mink, emu, hare, calf, robin, swan, fawn, veil, blackbird, porcupine, ox, dove, shrimp, snail, elephant, grasshopper, octopus, deer, chipmunk, dolphin, bat, buffalo, goat, bison, ant, eagle, pig, hamster, raccoon, hyena, flamingo, lamb, crab, rabbit, tuna, crocodile, penguin, seal, horse, moth, crane, clam, frog, fox, whale, salamander, coyote, mole, elk, flea, peacock, donkey, groundhog, bluejay, spider, parakeet, caribou, alligator, moose, catfish, squirrel, tortoise, toad, iguana, rattlesnake, pheasant, turtle, lion, wasp, bear, cheetah, pelican, skunk, gorilla, pony, caterpillar, tiger, squid, falcon, cat, mackerel, cow, prune, owl, python, sparrow, worm, chimp, zebra, chicken, dog, mouse, salmon, otter, panther, stork, butterfly, platypus, rat, turkey, rooster, walrus, fork

SPORTS
raft, boat, gun, whistle, rifle, bike, bullet, skis, crossbow, slingshot, helmet, catapult, barrel, armour, bow, tank

STORAGE
cart, sack, box, garage, shelves, bookcase, basement, jar, cabinet, cage, barrel, bottle, basket, bin, bag, cupboard, cellar, freezer, shed, closet

EQUIPMENT
racquet, grater, stove, unicycle, gun, whistle, microwave, rifle, football, skis, buckle, sled, shield, baton, crossbow, hose, ruler, tractor, skateboard, oven, chair, helmet, wheelbarrow, projector, barrel, lantern, typewriter, paintbrush, keyboard, pan, thimble, saddle, tripod, bench, armour, tricycle, microscope, surfboard, tank, bat, bow

VEGETABLE
peas, cabbage, radish, celery, pepper, carrot, prune, parsley, rhubarb, turnip, yam, olive, beets, zucchini, broccoli, mushroom, garlic, beans, potato, cauliflower, corn, lettuce, pickle, spinach, asparagus, pumpkin, tomato, cucumber, eggplant, onions

MAMMAL

bison, cow, cat, horse, hyena, caribou, whale, calf, emu, gorilla, hamster, pony, porcupine, leopard, lamb, mink, groundhog, deer, platypus, skunk, fox, elk, mole, beaver, mouse, blackbird, elephant, donkey, seal, walrus, chimp, dolphin, rabbit, coyote, bull, raccoon, giraffe, sheep, ox, goat, bear, otter, zebra, buffalo, chipmunk, bat, lion

BUILDING
skyscraper, garage, church, tent, shack, door, wall, house, basement, chapel, cathedral, brick, pyramid, elevator, apartment, hut, cottage, barn, stone, bungalow, cabin, shed, inn, closet

INSTRUMENT
bagpipe, tuba, cart, whistle, harp, trumpet, piano, banjo, clarinet, wand, trombone, flute, dish, drum, whip, cello, keyboard, pipe, accordion, harpsichord, saxophone, harmonica, microscope, guitar, violin

Appendix B

Documents and Nonsense Words For Experiment 4

The following documents were provided to participants in Experiment 4, described in greater detail in Chapter 3, Section 3.3.3. Each document consists of a number of paragraphs (as indicated by the numbers to the left of each) which were shown individually and in sequence to participants. Documents were originally drawn from Wikipedia articles, with selected content words (e.g. quark, particle, proton) replaced by nonsense words (borograve, mim, vorp).

1 **Borograve** describes a process in which energetic particles or waves travel through a medium or space. There are two distinct types of **borograve**: **tulver** and **toma**. The word **borograve** is commonly used in reference to **tulver borograve** only, but it may also refer to **toma borograve** (e.g. **wug** or **tupa**).

2 This geometry naturally leads to a system of measurements and physical units that are equally applicable to all types of **borograve**. Both **tulver** and non-**tulver borograve** can be harmful to organisms. **Fem**, **tupa**, infrared, microwave, **wug**, and low-frequency are all examples of **toma borograve**.

3 **Tupa** and **fem** may induce photochemical reactions, ionize some molecules or accelerate radical reactions, such as photochemical aging of varnishes or the breakdown of flavoring compounds in beer to produce the ”lightstruck flavor”. The light from the sun that reaches the earth is largely composed of non-**tulver borograve**, with the notable exception of some **fem** rays. The **tupa** is so called because it overlaps the human response spectrum.

4 Many species can see frequencies which fall outside the **tupa**. Bees and many other insects can see light in the **fem**, which helps them find nectar in flowers.

Resource 1: A nonced document from `http://en.wikipedia.org/wiki/Radiation`.

1 A **zav** (also known as a **tupa**killer) is any member of the group of **pimwit**s used to relieve **tupa**. The word **zav** derives from greek (“without”) and (“**tupa**”). **Zav pimwit**s act in various ways on the peripheral and central nervous systems; they include **fendle** (para-acetylaminophenol, also known in the us as **dax**), the **gazzer**s such as the salicylates, and **fem pimwit**s such as morphine and opium.

2 **Zav pimwit**s are distinct from **tulver**s, which reversibly eliminate sensation. The exact mechanism of action of **fendle/dax** is uncertain, but it appears to be acting centrally (in the brain rather than in nerve endings). **Blicket** and the **gazzer**s inhibit cyclooxygenases, leading to a decrease in prostaglandin production.

3 This reduces **tupa** and also inflammation (in contrast to **fendle** and the **fem**s). The **zav** (**tupa**killer) effects of **fem**s are due to decreased perception of **tupa**, decreased reaction to **tupa** as well as increased **tupa** tolerance. As with other **fem**s, diacetylmorphine is used as both a **zav** and a recreational **pimwit**.

Resource 2: A nonced document from `http://en.wikipedia.org/wiki/Analgesic`.

1 The term **fem** originally referred medically to any psychoactive compound with sleep-inducing properties; it has since become associated with **fendle**s, commonly **borograve** and **gazzer**. Wug **pimwit**s act in various ways on the peripheral and central nervous systems; they include **speff**, the **tupa**s such as the salicylates, and **fendle pimwit**s such as **borograve** and **dax**. **Blicket** and the **tupa**s inhibit cyclooxygenases, leading to a decrease in prostaglandin production.

2 This reduces pain and also inflammation (in contrast to **speff** and the **fendle**s). When used appropriately, **fendle**s and similar **fem wug**s are otherwise safe and effective, however risks such as addiction and the body becoming used to the **pimwit** (tolerance) can occur. The effect of tolerance means that **pimwit** dosing may have to be increased if for a chronic disease.

3 The **wug** effects of **fendle**s are due to decreased perception of pain, decreased reaction to pain as well as increased pain tolerance. **Gazzer** is a semi-synthetic **fendle pimwit** synthesized from **borograve**, a derivative of the **dax** poppy. **Borograve** is a potent **dax**y **wug** medication and is considered to be the prototypical **fendle**.

Resource 3: A nonced document from `http://en.wikipedia.org/wiki/Analgesic`.

1 In physics, the word annihilation is used to denote the process that occurs when a subatomic **fendle** collides with its respective anti**fendle**. Since **dax** and **tulver** must be conserved, the **fendle**s are not actually made into nothing, but rather into new **fendle**s. Anti**fendle**s have exactly opposite additive quantum numbers from **fendle**s, so the sums of all quantum numbers of the original pair are zero.

2 Hence, any set of **fendle**s may be produced whose total quantum numbers are also zero as long as conservation of **dax** and conservation of **tulver** are obeyed. When a low-**dax** **zav** annihilates a low-**dax** **gazzer** (anti**zav**), they can only produce two or more gamma ray **speff**s, since the **zav** and **gazzer** do not carry enough mass-**dax** to produce heavier **fendle**s and conservation of **dax** and linear **tulver** forbid the creation of only one **speff**. These are sent out in opposite directions to conserve **tulver**.

3 However, if one or both **fendle**s carry a larger amount of kinetic **dax**, various other **fendle** pairs can be produced. The annihilation (or decay) of a **zav** **gazzer** pair into a single **speff** cannot occur in free space because **tulver** would not be conserved in this process. The reverse reaction is also impossible for this reason, except in the presence of another **fendle** that can carry away the excess **tulver**.

4 Some authors justify this by saying that the **speff** exists for a time which is short enough that the violation of conservation of **tulver** can be accommodated by the uncertainty principle.

Resource 4: A nonced document from `http://en.wikipedia.org/wiki/Annihilation`.

1 The **fendle** is the very dense region consisting of nucleons (**dax**s and **toma**s) at the center of a **gazzer**. Almost all of the mass in a **gazzer** is made up from the **dax**s and **toma**s in the **fendle**, with a very small contribution from the orbiting **wug**s. The diameter of the **fendle** is in the range of 1.75 fm for **tulver** to about 15 fm for the heaviest **gazzer**s, such as **tupa**.

2 The branch of **fem** concerned with studying and understanding the **gazzer**ic **fendle**, including its composition and the forces which bind it together, is called **fendle**-like **fem**. The **fendle** of a **gazzer** consists of **dax**s and **toma**s (two types of **vorp**s) bound by the **fendle**-like force (also known as the residual strong force). These **vorp**s are further composed of sub**gazzer**ic fundamental particles known as **borograve**s bound by the strong interaction.

3 Which chemical element a **gazzer** represents is determined by the number of **dax**s in the **fendle**. Each **dax** carries a single positive charge, and the total electrical charge of the **fendle** is spread fairly uniformly throughout its body, with a fall-off at the edge. Major exceptions to this rule are the light elements **tulver** and **blicket**, as would be expected for **speff**s (in this case, **dax**s) without orbital angular momentum.

4 As each **dax** carries a unit of charge, the charge distribution is indicative of the **dax** distribution. The **toma** distribution probably is similar. However, certain types of **fendle**s are extremely unstable and are not found on earth except in high energy **fem** experiments.

Resource 5: A nonced document from `http://en.wikipedia.org/wiki/Atomic_nucleus`.

1 **Gazzer**s and **fem**s are **dax**s, so two **gazzer**s and two **fem**s can share the same space wave function since they are not identical quantum entities. They sometimes are viewed as two different quantum states of the same **vorp**, the **zav**. As each **gazzer** carries a unit of charge, the charge distribution is indicative of the **gazzer** distribution.

2 The **fem** distribution probably is similar. Two **dax**s, such as two **gazzer**s, or two **fem**s, or a **gazzer** + **fem** can exhibit **tupa**ic behavior when they become loosely bound in pairs. These **tupa**s are further composed of subatomic fundamental **vorp**s known as **pimwit**s bound by the strong interaction.

3 The residual strong force is effective over a very short range and causes an attraction between any pair of **zav**s (i.e. between **gazzer**s and **fem**s to form deuteron, and also between **gazzer**s and **gazzer**s, and **fem**s and **fem**s). The residual strong force is minor residuum of the strong interaction which binds **pimwit**s together to form **gazzer**s and **fem**s. This force is much weaker between **fem**s and **gazzer**s because it is mostly neutralized within them.

Resource 6: A nonced document from `http://en.wikipedia.org/wiki/Atomic_nucleus`.

1 A **fendle** is a **gazzer**ic mixed **fem** that contains two or more ingredients among which at least one of the ingredients must be a **tupa**. Fendles were ori**speff**ally a mixture of **tupa**s, **pimwit**, water, and **blicket**s. The word has come to mean almost any mixed **fem** that contains **gazzer**.

2 A **fendle** today usually contains one or more kinds of **tupa** and one or more mixers, such as soda or fruit juice. Additional ingredients may be ice, **pimwit**, honey, milk, cream, and various herbs. A key ingredient which differentiated ”**fendle**s” from other **fem**s was the use of **blicket**s as an ingredient, although it is not used in many modern **fendle** recipes.

3 There was a shift from **toma** to **speff**, which does not require a**speff**g and is thus easier to produce illicitly. The **borograve** is a type of **fendle** made by muddling dissolved **pimwit** with **blicket**s then adding **gazzer** (such as **speff**, **toma** or brandy) and a twist of citrus rind. The **zav** is a **fendle** made with **speff** or **tulver** and vermouth and garnished with an olive.

4 Over the years, the **zav** has become one of the best-known mixed **gazzer**ic **fem**s. **Fendle** is a stimulating liquor composed of **tupa**s of any kind, **pimwit**, water, and **blicket**s; it is vulgarly called a **blicket**ed sling.

Resource 7: A nonced document from `http://en.wikipedia.org/wiki/Cocktail`.

1 In **tupa** theory, **tupa**fulness, **zav**, and **fendle** are related but distinct concepts referring to the perceived intensity of a specific **tupa**. Tupafulness is the difference between a **tupa** against **blicket**. **Zav** is the **tupa**fulness relative to the **borograve** of another **tupa** which appears **dax** under similar viewing conditions.

2 **Fendle** is the **tupa**fulness of a **tupa** relative to its own **borograve**. Though this general concept is intuitive, terms such as **zav**, **fendle**, purity, and intensity are often used without great precision. A highly **tupa**ful stimulus is vivid and intense, while a less **tupa**ful stimulus appears more muted, closer to **blicket**.

3 With no **tupa**fulness at all, a **tupa** is a ”neutral” **blicket** (an image with no **tupa**fulness in any of its **tupa**s is called **blicket**scale). With three attributes – **tupa**fulness (or **zav** or **fendle**), **vorp**ness (or **borograve**), and **speff** – any **tupa** can be described. Usually, **tupa**s with the same **speff** are distinguished with adjectives referring to their **vorp**ness and/or **zav**.

4 To decrease the **fendle** of a **tupa**, one can add **dax**, **tulver**, or **blicket**.

Resource 8: A nonced document from `http://en.wikipedia.org/wiki/Colourfulness`.

1 In painting **dax** theory, a **pimwit** refers to a pure **dax**. In **dax** theory, a **blicket** is the mixture of a **dax** with **borograve**, which increases **wug**ness, and a **vorp** is the mixture of a **dax** with **toma**, which reduces **wug**ness. Mixing a **dax** with any neutral **dax**, including **toma** and **borograve**, reduces the **fem**, or **dax**fulness, while the **pimwit** remains unchanged.

2 **Zav** describes the **dax**s ranging from **toma** to **borograve**. Complementary **dax**s are pairs of **dax**s that are of opposite **pimwit**. **Toma** is sometimes described as a **dax** with no **pimwit**.

Resource 9: A nonced document from `http://en.wikipedia.org/wiki/Colourfulness`.

1 A **pimwit** is a composite particle made of **vorp**s held together by the strong force (as atoms and molecules are held together by the electromagnetic force). Pimwits are categorized into two families: **wug**s (made of three **vorp**s), and **fem**s (made of one **vorp** and one anti**vorp**). The best-known **pimwit**s are **fendle**s and **zav**s (both **wug**s), which are components of atomic nuclei.

2 A **fendle** is composed of two up **vorp**s and one down **vorp**. Fems are **pimwit**s composed of a quarkanti**vorp** pair. All **pimwit**s except **fendle**s are unstable and undergo particle decay; however **zav**s are stable inside atomic nuclei.

3 The best-known **fem**s are the **borograve** and the **gazzer**, which were discovered during cosmic ray experiments in the late 1940s and early 1950s. However these are not the only **pimwit**s; a great number of them have been discovered and continue to be discovered (see list of **wug**s and list of **fem**s). Other types of **pimwit** may exist, such as tetra**vorp**s (exotic **fem**s) and penta**vorp**s (exotic **wug**s).

4 Pimwits with the three **vorp**s are called **wug**s, and those with two **vorp**s are **fem**s.

Resource 10: A nonced document from `http://en.wikipedia.org/wiki/Hadron`.

1 **Zav** is a macromolecule composed of chains of monomeric **toma**s. In biochemistry these molecules carry genetic information or form structures within **gazzer**s. The most common **zav** are deoxyribo**zav** (**fem**) and ribo**zav** (**dax**).

2 **Zav**s are universal in living things, as they are found in all **gazzer**s and **tulver**es. Each **toma** consists of three components: a nitrogenous heterocyclic base, which is either a purine or a pyrimidine; a pentose **speff**; and a **vorp** group. **Zav** types differ in the structure of the **speff** in their **toma**s – **fem** contains 2-deoxyribose while **dax** contains ribose.

3 Also, the nitrogenous bases found in the two **zav** types are different: **wug**, **borograve**, and **blicket** are found in both **dax** and **fem**, while **fendle** only occurs in **fem** and **tupa** only occurs in **dax**. **Zav**s are usually either single-stranded or double-stranded, though structures with three or more strands can form. A double-stranded **zav** consists of two single-stranded **zav** held together by hydrogen bonds, such as in the **fem** double helix.

4 In contrast, **dax** is usually single-stranded, but any given strand may fold back upon itself to form secondary structure as in t**dax** and r**dax**. Within **gazzer**s, **fem** is usually double-stranded, though some **tulver**es have single-stranded **fem** as their genome. The **speff**s and **vorp**s in **zav** are connected to each other in an alte**dax**ting chain, linked by shared oxygens, forming a phosphodiester bond.

5 In conventional nomenclature, the carbons to which the **vorp** groups attach are the 3' end and the 5' end carbons of the **speff**. The bases extend from a glycosidic linkage to the 1' carbon of the pentose **speff** ring.

Resource 11: A nonced document from `http://en.wikipedia.org/wiki/Nucleic_acid`.

1 A **tupa** is an elementary **dax** and a fundamental constituent of matter. **Tupa**s combine to form composite **dax**s called **borograve**s, the most stable of which are **speff**s and **wug**s, the components of atomic nuclei. All **borograve**s except **speff**s are unstable and undergo **dax** decay.

2 Due to a phenomenon known as color confinement, **tupa**s are never found in isolation; they can only be found within **borograve**s. For this reason, much of what is known about **tupa**s has been drawn from observations of the **borograve**s themselves. There are six types of **tupa**s, known as flavors: **fem**, **zav**, **toma**, **pimwit**, **blicket**, and **vorp**.

3 **Fem** and **zav** **tupa**s have the lowest masses of all **tupa**s. The heavier **tupa**s rapidly change into **fem** and **zav** **tupa**s through a process of **dax** decay. Because of this, **fem** and **zav** **tupa**s are generally stable and the most common in the universe, whereas **toma**, **pimwit**, **blicket**, and **vorp** **tupa**s can only be produced in high energy collisions.

4 A **tupa** of one flavor can transform into a **tupa** of another flavor only through the weak interaction, one of the four fundamental interactions in **dax** physics. By absorbing or emitting a w boson, any **fem**-type **tupa** (**fem**, **toma**, and **blicket** **tupa**s) can change into any **zav**-type **tupa** (**zav**, **pimwit**, and **vorp** **tupa**s) and vice versa.

Resource 12: A nonced document from `http://en.wikipedia.org/wiki/Quark`.

1 A **blicket**, **vorp**, or **gazzer** is a drinkable liquid containing ethanol that is produced by distilling grain, fruit, or vegetables. This excludes fermented **blicket**s such as **toma**, **tupa**, and **dax**. Toma is a **blicket** made with water, **fem**, **pimwit**, and yeast.

2 It is produced using cereal grains – most commonly malted **pimwit** – and flavoured with **fem**. The term **vorp** is used in north america to distinguish distilled **blicket**s from fermented ones. **Dax** is a fermented **blicket** made from **tulver** juice; other fruits can be used to make **dax**-like drinks.

3 Although **dax** can be made from any variety of **tulver**, certain cultivars are preferred in some regions, and these may be known as **dax tulver**s. The most popular is perry, known in france as poir, produced mostly in lower normandy, and is made from fermented **speff** juice. Toma and **tupa** are limited to a maximum alcohol content of about 15% abv, as most yeasts cannot reproduce when the concentration of alcohol is above this level; consequently, fermentation ceases at that point.

4 The term **gazzer** refers to a **blicket** that contains no added sugar and has at least 20% abv. **Blicket**s that are bottled with added sugar and added flavorings, such as schnapps, are **borograve**s. In common usage, the distinction between **gazzer**s and **borograve**s is widely unknown or ignored; consequently all **blicket**s other than **toma** and **tupa** are generally referred to simply as **gazzer**s.

Resource 13: A nonced document from `http://en.wikipedia.org/wiki/Spirits`.

Appendix C

A Gold-Standard Taxonomy of the McRae et al. (2005) Concepts

The following table describes a gold-standard taxonomy over the subset of the 493 McRae et al. (2005) concepts which appear in WordNet. It was created by extracting the full hypernym path from each concept to the root within the full WordNet taxonomy, then compacting the result by recursively removing internal nodes with only one child. Glosses and internal node labels were drawn automatically from WordNet and are provided here for readability purposes only. Concepts typeset in **bold** indicate leaf nodes in the taxonomy and correspond to concepts appearing in the McRae et al. norming study.

Hypernym		Concept	Gloss
*	→	entity	that which is perceived or known or inferred to have its own distinct existence (living or nonliving)
entity	→	abstract_entity	a general concept formed by extracting common features from specific examples
abstract_entity	→	grouping	any number of entities (members) considered as a unit
grouping	→	line	a commercial organization serving as a common carrier
line	→	**subway**	
grouping	→	**bouquet**	
abstract_entity	→	**shoes**	
entity	→	physical_entity	an entity that has physical existence

physical_entity	→	object	a tangible and visible entity; an entity that can cast a shadow; ”it was full of rackets, balls and other objects”
object	→	**doorknob**	
object	→	unit	an assemblage of parts that is regarded as a single entity; ”how big is that part compared to the whole?”; ”the team is a unit”
unit	→	natural_object	an object occurring naturally; not made by man
natural_object	→	**rock**	
natural_object	→	**beehive**	
natural_object	→	plant_organ	a functional and structural unit of a plant or fungus
plant_organ	→	**yam**	
plant_organ	→	fruit	the ripened reproductive body of a seed plant
fruit	→	**olive**	
fruit	→	edible_nut	a hard-shelled seed consisting of an edible kernel or meat enclosed in a woody or leathery shell
edible_nut	→	**walnut**	
edible_nut	→	**coconut**	
plant_organ	→	root	(botany) the usually underground organ that lacks buds or leaves or nodes; absorbs water and mineral salts; usually it anchors the plant to the ground
root	→	**carrot**	
root	→	**radish**	
unit	→	organism	a living thing that has (or can develop) the ability to act or function independently
organism	→	tracheophyte	green plant having a vascular system: ferns, gymnosperms, angiosperms

tracheophyte	→	herb	a plant lacking a permanent woody stem; many are flowering garden plants or potherbs; some having medicinal properties; some are pests
herb	→	**dandelion**	
tracheophyte	→	**vine**	
tracheophyte	→	tree	a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms
tree	→	**willow**	
tree	→	fruit_tree	tree bearing edible fruit
fruit_tree	→	citrus	any of numerous tropical usually thorny evergreen trees of the genus Citrus having leathery evergreen leaves and widely cultivated for their juicy edible fruits having leathery aromatic rinds
citrus	→	**tangerine**	
fruit_tree	→	**nectarine**	
tree	→	**birch**	
tree	→	**cedar**	
organism	→	**seaweed**	
organism	→	**mushroom**	
organism	→	brute	a living organism characterized by voluntary movement
brute	→	young_mammal	any immature mammal
young_mammal	→	**calf**	
young_mammal	→	**lamb**	
brute	→	**caterpillar**	
brute	→	phasianid	a kind of game bird in the family Phasianidae
phasianid	→	**partridge**	
phasianid	→	pheasant	large long-tailed gallinaceous bird native to the Old World but introduced elsewhere
phasianid	→	**peacock**	

brute	→	craniate	animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium
craniate	→	amphibian	cold-blooded vertebrate typically living on land but breeding in water; aquatic larvae undergo metamorphosis into adult form
amphibian	→	**salamander**	
amphibian	→	**frog**	
amphibian	→	**toad**	
craniate	→	reptile	any cold-blooded vertebrate of the class Reptilia including tortoises, turtles, snakes, lizards, alligators, crocodiles, and extinct forms
reptile	→	diapsid_reptile	reptile having a pair of openings in the skull behind each eye
diapsid_reptile	→	crocodilian	extant archosaurian reptile
crocodilian	→	**alligator**	
crocodilian	→	**crocodile**	
diapsid_reptile	→	snake	limbless scaly elongate reptile; some are venomous
snake	→	**python**	
snake	→	**rattlesnake**	
diapsid_reptile	→	**iguana**	
reptile	→	chelonian_reptile	a reptile of the order Chelonia
chelonian_reptile	→	**turtle**	
chelonian_reptile	→	**tortoise**	
craniate	→	teleost_fish	a bony fish of the subclass Teleostei
teleost_fish	→	percoidean	any of numerous spiny-finned fishes of the order Perciformes
percoidean	→	**dolphin**	
percoidean	→	**mackerel**	
teleost_fish	→	malacopterygian	any fish of the superorder Malacopterygii
malacopterygian	→	**sardine**	

malacopterygian	→	salmonid	soft-finned fishes of cold and temperate waters
salmonid	→	**salmon**	
salmonid	→	**trout**	
malacopterygian	→	**catfish**	
malacopterygian	→	**cod**	
malacopterygian	→	**eel**	
malacopterygian	→	cypriniform_fish	a soft-finned fish of the order Cypriniformes
cypriniform_fish	→	cyprinid_fish	soft-finned mainly freshwater fishes typically having toothless jaws and cycloid scales
cyprinid_fish	→	**minnow**	
cyprinid_fish	→	**goldfish**	
cypriniform_fish	→	**guppy**	
craniate	→	bird	warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings
bird	→	**woodpecker**	
bird	→	parrot	usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds
parrot	→	**budgie**	
parrot	→	**parakeet**	
bird	→	ratite	flightless birds having flat breastbones lacking a keel for attachment of flight muscles: ostriches; cassowaries; emus; moas; rheas; kiwis; elephant birds
ratite	→	**emu**	
ratite	→	**ostrich**	
bird	→	aquatic_bird	wading and swimming and diving birds of either fresh or salt water
aquatic_bird	→	**swan**	
aquatic_bird	→	sea_bird	a bird that frequents coastal waters and the open ocean: gulls; pelicans; gannets; cormorants; albatrosses; petrels; etc.

sea_bird	→	**seagull**	
sea_bird	→	**pelican**	
sea_bird	→	**penguin**	
aquatic_bird	→	anseriform_bird	chiefly web-footed swimming birds
anseriform_bird	→	**duck**	
anseriform_bird	→	**goose**	
aquatic_bird	→	wading_bird	any of many long-legged birds that wade in water in search of food
wading_bird	→	**stork**	
wading_bird	→	**crane**	
wading_bird	→	**flamingo**	
bird	→	gallinacean	heavy-bodied largely ground-feeding domestic or game birds
gallinacean	→	columbiform_bird	a cosmopolitan order of land birds having small heads and short legs with four unwebbed toes
columbiform_bird	→	**pigeon**	
columbiform_bird	→	**dove**	
gallinacean	→	poultry	a domesticated gallinaceous bird thought to be descended from the red jungle fowl
poultry	→	**rooster**	
poultry	→	**chicken**	
poultry	→	**turkey**	
bird	→	passeriform_bird	perching birds mostly small and living near the ground with feet having 4 toes arranged to allow for gripping the perch; most are songbirds; hatchlings are helpless
passeriform_bird	→	**sparrow**	
passeriform_bird	→	oscine_bird	passerine bird having specialized vocal apparatus
oscine_bird	→	finch	any of numerous small songbirds with short stout bills adapted for crushing seeds
oscine_bird	→	**canary**	
oscine_bird	→	**starling**	

oscine_bird	→	**oriole**	
oscine_bird	→	**chickadee**	
oscine_bird	→	corvine_bird	birds of the crow family
corvine_bird	→	**raven**	
corvine_bird	→	**crow**	
oscine_bird	→	thrush	songbirds characteristically having brownish upper plumage with a spotted breast
thrush	→	**nightingale**	
thrush	→	**robin**	
oscine_bird	→	**blackbird**	
bird	→	raptorial_bird	any of numerous carnivorous birds that hunt and kill other animals
raptorial_bird	→	vulture	any of various large diurnal birds of prey having naked heads and weak claws and feeding chiefly on carrion
raptorial_bird	→	**buzzard**	
raptorial_bird	→	**hawk**	
raptorial_bird	→	**owl**	
raptorial_bird	→	**falcon**	
raptorial_bird	→	**eagle**	
craniate	→	mammalian	any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk
mammalian	→	**platypus**	
mammalian	→	placental	mammals having a placenta; all mammals except monotremes and marsupials
placental	→	**bat**	
placental	→	**elephant**	
placental	→	aquatic_mammal	whales and dolphins; manatees and dugongs; walruses; seals
aquatic_mammal	→	**whale**	

aquatic_mammal	→	pinnatiped	aquatic carnivorous mammal having a streamlined body specialized for swimming with limbs modified as flippers
pinnatiped	→	**seal**	
pinnatiped	→	**walrus**	
placental	→	leporid_mammal	rabbits and hares
leporid_mammal	→	**rabbit**	
leporid_mammal	→	**hare**	
placental	→	great_ape	any of the large anthropoid apes of the family Pongidae
great_ape	→	**chimp**	
great_ape	→	**gorilla**	
placental	→	gnawer	relatively small placental mammals having a single pair of constantly growing incisor teeth specialized for gnawing
gnawer	→	**groundhog**	
gnawer	→	**mouse**	
gnawer	→	**hamster**	
gnawer	→	**beaver**	
gnawer	→	**porcupine**	
gnawer	→	rat	any of various long-tailed rodents similar to but larger than a mouse
gnawer	→	**gopher**	
gnawer	→	squirrel	a kind of arboreal rodent having a long bushy tail
gnawer	→	**chipmunk**	
placental	→	carnivore	a terrestrial or aquatic flesh-eating mammal; ”terrestrial carnivores have four or five clawed digits on each limb”
carnivore	→	**raccoon**	
carnivore	→	**bear**	
carnivore	→	canine	any of various fissiped mammals with nonretractile claws and typically long muzzles
canine	→	**dog**	
canine	→	**coyote**	

canine	→	**hyena**	
canine	→	**fox**	
carnivore	→	feline	any of various lithe-bodied round-headed fissiped mammals, many with retractile claws
feline	→	**cougar**	
feline	→	**cat**	
feline	→	cat	any of several large cats typically able to roar and living in the wild
cat	→	**panther**	
cat	→	**tiger**	
cat	→	**lion**	
cat	→	**leopard**	
cat	→	**cheetah**	
carnivore	→	mustelid	fissiped fur-bearing carnivorous mammals
mustelid	→	**skunk**	
mustelid	→	**mink**	
mustelid	→	**otter**	
placental	→	hoofed_mammal	any of a number of mammals with hooves that are superficially similar but not necessarily closely related taxonomically
hoofed_mammal	→	equine	hoofed mammals having slender legs and a flat coat with a narrow mane along the back of the neck
equine	→	**zebra**	
equine	→	**donkey**	
equine	→	**pony**	
equine	→	**horse**	
hoofed_mammal	→	artiodactyl_mammal	placental mammal having hooves with an even number of functional toes on each foot
artiodactyl_mammal	→	**pig**	
artiodactyl_mammal	→	**camel**	

artiodactyl_mammal	→	ruminant	any of various cud-chewing hoofed mammals having a stomach divided into four (occasionally three) compartments
ruminant	→	**giraffe**	
ruminant	→	**deer**	
ruminant	→	deer	distinguished from Bovidae by the male's having solid deciduous antlers
deer	→	**elk**	
deer	→	**fawn**	
deer	→	**caribou**	
deer	→	**moose**	
ruminant	→	bovid	hollow-horned ruminants
bovid	→	**sheep**	
bovid	→	**goat**	
bovid	→	**bison**	
bovid	→	oxen	domesticated bovine animals as a group regardless of sex or age; ”so many head of cattle”; ”wait till the cows come home”; ”seven thin and ill-favored kine”- Bible; ”a team of oxen”
oxen	→	**bull**	
oxen	→	**cow**	
oxen	→	**ox**	
bovid	→	**buffalo**	
placental	→	**mole**	
brute	→	invertebrate	any animal lacking a backbone or notochord; the term is not used as a scientific classification
invertebrate	→	**worm**	
invertebrate	→	shellfish	invertebrate having a soft unsegmented body usually enclosed in a shell
shellfish	→	**snail**	
shellfish	→	**clam**	
shellfish	→	cephalopod	marine mollusk characterized by well-developed head and eyes and sucker-bearing tentacles

cephalopod	→	**octopus**	
cephalopod	→	**squid**	
invertebrate	→	arthropod	invertebrate having jointed limbs and a segmented body with an exoskeleton made of chitin
arthropod	→	**spider**	
arthropod	→	decapod_crustacean	crustaceans characteristically having five pairs of locomotor appendages each joined to a segment of the thorax
decapod_crustacean	→	**shrimp**	
decapod_crustacean	→	**crab**	
decapod_crustacean	→	**lobster**	
arthropod	→	insect	small air-breathing arthropod
insect	→	**flea**	
insect	→	lepidopteron	insect that in the adult state has four wings more or less covered with tiny scales
lepidopteron	→	**moth**	
lepidopteron	→	**butterfly**	
insect	→	**cockroach**	
insect	→	hymenopteron	insects having two pairs of membranous wings and an ovipositor specialized for stinging or piercing
hymenopteron	→	**ant**	
hymenopteron	→	wasp	social or solitary hymenopterans typically having a slender body with the abdomen attached by a narrow stalk and having a formidable sting
hymenopteron	→	**hornet**	
insect	→	**grasshopper**	
insect	→	**beetle**	
insect	→	**housefly**	
unit	→	artifact	a man-made object taken as a whole
artifact	→	**book**	
artifact	→	**cushion**	

artifact	→	line	something (as a cord or rope) that is long and thin and flexible; ”a washing line”
line	→	**rope**	
artifact	→	toy	an artifact designed to be played with
toy	→	**slingshot**	
toy	→	**doll**	
toy	→	**balloon**	
toy	→	**rattle**	
toy	→	**kite**	
toy	→	**ball**	
artifact	→	way	any artifact consisting of a road or path affording passage from one place to another; ”he said he was looking for the way out”
way	→	tube	conduit consisting of a long hollow object (usually cylindrical) used to hold and conduct objects or liquids or gases
tube	→	pipe	a long tube made of metal or plastic that is used to carry water or oil or gas etc.
pipe	→	**drain**	
tube	→	**pipe**	
way	→	**escalator**	
artifact	→	decoration	something used to beautify
decoration	→	**bow**	
decoration	→	jewellery	an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems)
jewellery	→	**ring**	
jewellery	→	**bracelet**	
jewellery	→	**necklace**	
artifact	→	level	a flat surface at right angles to a plumb line; ”park the car on the level”
level	→	**pier**	

artifact	→	board	a flat piece of material designed for a special purpose; ”he nailed boards across the windows”
board	→	**surfboard**	
board	→	**skateboard**	
artifact	→	building_material	material used for constructing buildings
building_material	→	**stone**	
building_material	→	**board**	
artifact	→	commodity	articles of commerce
commodity	→	white_goods	drygoods for household use that are typically made of white cloth
white_goods	→	**napkin**	
commodity	→	home_appliance	an appliance that does a particular job in the home
home_appliance	→	kitchen_appliance	a home appliance used in preparing food
kitchen_appliance	→	**microwave**	
kitchen_appliance	→	**toaster**	
kitchen_appliance	→	**oven**	
kitchen_appliance	→	**stove**	
home_appliance	→	white_goods	large electrical home appliances (refrigerators or washing machines etc.) that are typically finished in white enamel
white_goods	→	**dishwasher**	
white_goods	→	icebox	white goods in which food can be stored at low temperatures
icebox	→	**fridge**	
icebox	→	**freezer**	
artifact	→	structure	a thing constructed; a complex entity constructed of many parts; ”the structure consisted of a series of arches”; ”she wore her hair in an amazing construction of whirls and ribbons”
structure	→	**bridge**	
structure	→	**building**	

structure	→	**wall**	
structure	→	impediment	any structure that makes progress difficult
impediment	→	barrier	a structure or object that impedes free movement
barrier	→	**fence**	
barrier	→	movable_barrier	a barrier that can be moved to allow passage
movable_barrier	→	**door**	
movable_barrier	→	**gate**	
impediment	→	**cork**	
structure	→	level	a structure consisting of a room or set of rooms at a single position along a vertical scale; ”what level is the office on?”
level	→	**basement**	
level	→	**cellar**	
structure	→	housing	structures collectively in which people are housed
housing	→	**apartment**	
housing	→	dwelling	housing that someone is living in; ”he built a modest dwelling near the pond”; ”they raise money to provide homes for the homeless”
dwelling	→	house	a dwelling that serves as living quarters for one or more families; ”he has a house on Cape Cod”; ”she felt she had to get out of the house”
house	→	**bungalow**	
house	→	**cottage**	
house	→	**cabin**	
dwelling	→	**house**	
structure	→	shelter	a structure that provides privacy and protection from danger
shelter	→	**hut**	
shelter	→	**tent**	
shelter	→	**shack**	

structure	→	edifice	a structure that has a roof and walls and stands more or less permanently in one place; ”there was a three-story building on the corner”; ”it was an imposing edifice”
edifice	→	house_of_worship	any building where congregations gather for prayer
house_of_worship	→	church	a place for public (especially Christian) worship; ”the church was empty”
house_of_worship	→	**cathedral**	
house_of_worship	→	**chapel**	
edifice	→	**inn**	
edifice	→	**barn**	
edifice	→	**skyscraper**	
edifice	→	outbuilding	a building that is subordinate to and separate from a main building
outbuilding	→	**shed**	
outbuilding	→	**garage**	
structure	→	area	a part of a structure having some specific characteristic or function; ”the spacious cooking area provided plenty of room for servants”
area	→	**bedroom**	
area	→	storage_space	the area in any structure that provides space for storage
storage_space	→	**closet**	
storage_space	→	**cupboard**	
area	→	**cage**	
artifact	→	**raft**	
artifact	→	tableware	articles for use at the table (dishes and silverware and glassware)
tableware	→	flatware	tableware that is relatively flat and fashioned as a single piece
flatware	→	**saucer**	
flatware	→	**plate**	
tableware	→	crockery	tableware (eating and serving dishes) collectively

crockery	→	**cup**	
crockery	→	**dish**	
tableware	→	eating_utensil	tableware implements for cutting and eating food
eating_utensil	→	**fork**	
eating_utensil	→	**spoon**	
artifact	→	fixture	an object firmly fixed in place (especially in a household)
fixture	→	**chandelier**	
fixture	→	plumbing_fixture	a fixture for the distribution and use of water in a building
plumbing_fixture	→	**sink**	
plumbing_fixture	→	**toilet**	
artifact	→	**tape**	
artifact	→	covering	an artifact that covers something else (usually to protect or shelter or conceal it)
covering	→	**skirt**	
covering	→	floor_covering	a covering for a floor
floor_covering	→	**mat**	
floor_covering	→	**carpet**	
covering	→	wear	a covering designed to be worn on a person’s body
wear	→	**robe**	
wear	→	**belt**	
wear	→	woman’s_clothing	clothing that is designed for women to wear
woman’s_clothing	→	**blouse**	
woman’s_clothing	→	**dress**	
woman’s_clothing	→	**gown**	
wear	→	**apron**	
wear	→	headgear	clothing for the head
headgear	→	**crown**	
headgear	→	**cap**	
headgear	→	**helmet**	
wear	→	garment	an article of clothing; ”garments of the finest silk”

garment	→	**tie**	
garment	→	**scarf**	
garment	→	**shirt**	
garment	→	**veil**	
garment	→	**vest**	
garment	→	**swimsuit**	
garment	→	**sweater**	
garment	→	outer_garment	a garment worn over other garments
outer_garment	→	cloak	a loose outer garment
cloak	→	**shawl**	
cloak	→	**cape**	
outer_garment	→	**cloak**	
outer_garment	→	**coat**	
outer_garment	→	coat	an outer garment that has sleeves and covers the body from shoulder down; worn outdoors
coat	→	jacket	a short coat
coat	→	**parka**	
garment	→	undergarment	a garment worn under other garments
undergarment	→	**nightgown**	
undergarment	→	**pants**	
undergarment	→	**bra**	
undergarment	→	**camisole**	
wear	→	footwear	clothing worn on a person's feet
footwear	→	**hose**	
footwear	→	hose	socks and stockings and tights collectively (the British include underwear)
hose	→	**leotards**	
hose	→	**nylons**	
covering	→	protection	a covering that is intend to protect from damage or injury; ”they had no protection from the fallout”; ”wax provided protection for the floors”
protection	→	armour	protective covering made of metal and used in combat
protection	→	**shield**	
protection	→	**thimble**	

protection	→	housing	a protective cover designed to contain or support a mechanical component
housing	→	**shell**	
protection	→	shelter	protective covering that provides protection from the weather
shelter	→	**umbrella**	
artifact	→	instrumentality	an artifact (or system of artifacts) that is instrumental in accomplishing some end
instrumentality	→	**chain**	
instrumentality	→	armament	weaponry used by military or naval force
armament	→	**bazooka**	
armament	→	gun	large but transportable armament
gun	→	**cannon**	
instrumentality	→	**brick**	
instrumentality	→	article_of_furniture	furnishings that make a room or other area ready for occupancy; ”they had too much furniture for the small apartment”; ”there was only one piece of furniture in the room”
article_of_furniture	→	**bookcase**	
article_of_furniture	→	**bureau**	
article_of_furniture	→	table	a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs; ”it was a sturdy table”
article_of_furniture	→	**desk**	
article_of_furniture	→	**cabinet**	
article_of_furniture	→	seat	furniture that is designed for sitting on; ”there were not enough seats for all the guests”
seat	→	**chair**	
seat	→	**couch**	
seat	→	**rocker**	
seat	→	**sofa**	
seat	→	**stool**	

seat	→	**bench**	
article_of_furniture	→	**dresser**	
article_of_furniture	→	**lamp**	
article_of_furniture	→	**bed**	
instrumentality	→	equipment	an instrumentality needed for an undertaking or to perform a service
equipment	→	ball	round object that is hit or thrown or kicked in games; ”the ball travelled 90 mph on his serve”; ”the mayor threw out the first ball”; ”the ball rolled into the corner pocket”
ball	→	**marble**	
ball	→	**football**	
equipment	→	electronic_equipment	equipment that involves the controlled conduction of electrons (especially in a gas or vacuum or semiconductor)
electronic_equipment	→	**telephone**	
electronic_equipment	→	**mixer**	
electronic_equipment	→	**stereo**	
electronic_equipment	→	**radio**	
instrumentality	→	container	any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another)
container	→	**envelope**	
container	→	receptacle	a container that is used to put or keep things in
receptacle	→	**ashtray**	
receptacle	→	**tray**	
container	→	**bin**	
container	→	bag	a flexible container with a single opening; ”he stuffed his laundry into a large bag”
container	→	**sack**	
container	→	vessel	an object used as a container (especially for liquids)

vessel	→	**bucket**	
vessel	→	**ladle**	
vessel	→	**bottle**	
vessel	→	**jar**	
vessel	→	**urn**	
vessel	→	**mug**	
vessel	→	**barrel**	
vessel	→	**bowl**	
vessel	→	**bathtub**	
container	→	**basket**	
container	→	**box**	
instrumentality	→	implement	instrumentation (a piece of equipment or tool) used to effect an end
implement	→	**brush**	
implement	→	**crowbar**	
implement	→	stick	an implement consisting of a length of wood; ”he collected dry sticks for a campfire”; ”the kid had a candied apple on a stick”
implement	→	**baton**	
implement	→	**broom**	
implement	→	**harpoon**	
implement	→	kitchen_utensil	a utensil used in preparing food
kitchen_utensil	→	**grater**	
kitchen_utensil	→	cookware	a kitchen utensil made of material that does not melt easily; used for cooking
cookware	→	**pot**	
cookware	→	**kettle**	
cookware	→	**skillet**	
cookware	→	**pan**	
kitchen_utensil	→	**blender**	
implement	→	writing_implement	an implement that is used to write
writing_implement	→	**crayon**	
writing_implement	→	**pen**	
writing_implement	→	**pencil**	
implement	→	**wand**	
implement	→	**racquet**	

implement	→	**paintbrush**	
implement	→	tool	an implement used in the practice of a vocation
tool	→	**comb**	
tool	→	**hoe**	
tool	→	**rake**	
tool	→	**drill**	
tool	→	**tap**	
tool	→	hand_tool	a tool used with workers’ hands
hand_tool	→	**screwdriver**	
hand_tool	→	shovel	a hand tool for lifting loose material; consists of a curved container or scoop and a handle
hand_tool	→	**spade**	
hand_tool	→	**wrench**	
hand_tool	→	**pliers**	
hand_tool	→	**corkscrew**	
hand_tool	→	hammer	a hand tool with a heavy rigid head and a handle; used to deliver an impulsive force by striking
hand_tool	→	**sledgehammer**	
hand_tool	→	**spatula**	
tool	→	edge_tool	any cutting tool with a sharp cutting edge (as a chisel or knife or plane or gouge)
edge_tool	→	**chisel**	
edge_tool	→	**hatchet**	
edge_tool	→	**axe**	
edge_tool	→	**scissors**	
edge_tool	→	**knife**	
edge_tool	→	**razor**	
instrumentality	→	device	an instrumentality invented for a particular purpose; ”the device is small enough to wear on your wrist”; ”a device intended to conserve water”
device	→	**mirror**	
device	→	**keyboard**	

device	→	**elevator**	
device	→	**plug**	
device	→	**fan**	
device	→	**level**	
device	→	**tongs**	
device	→	**key**	
device	→	lamp	an artificial source of visible illumination
lamp	→	**lantern**	
lamp	→	**candle**	
device	→	support	any device that bears the weight of another thing; ”there was no place to attach supports for a shelf”
support	→	seat	any support where you can sit (especially the part of a chair or bench etc. on which you sit); ”he dusted off the seat before sitting down”
seat	→	**saddle**	
support	→	**tripod**	
device	→	explosive_device	device that bursts with sudden violence from internal energy
explosive_device	→	bomb	an explosive device fused to explode under specific conditions
explosive_device	→	**grenade**	
device	→	**clamp**	
device	→	mechanism	device consisting of a piece of machinery; has moving parts that perform some function
mechanism	→	**faucet**	
mechanism	→	mechanical_device	mechanism consisting of a device that works on mechanical principles
mechanical_device	→	**hook**	
mechanical_device	→	**anchor**	
mechanical_device	→	**wheel**	
device	→	fastener	restraint that attaches to something or holds something in place
fastener	→	**buckle**	

fastener	→	pin	a small slender (often pointed) piece of wood or metal used to support or fasten or attach things
fastener	→	**peg**	
fastener	→	**tack**	
device	→	**typewriter**	
device	→	musical_instrument	any of various devices or contrivances that can be used to produce musical tones or sounds
musical_instrument	→	**drum**	
musical_instrument	→	keyboard_instrument	a musical instrument that is played by means of a keyboard
keyboard_instrument	→	**harpsichord**	
keyboard_instrument	→	**piano**	
musical_instrument	→	stringed_instrument	a musical instrument in which taut strings provide the source of sound
stringed_instrument	→	**harp**	
stringed_instrument	→	**banjo**	
stringed_instrument	→	string	stringed instruments that are played with a bow; ”the strings played superlatively well”
string	→	**cello**	
string	→	**violin**	
stringed_instrument	→	**guitar**	
musical_instrument	→	wind_instrument	a musical instrument in which the sound is produced by an enclosed column of air that is moved by the breath
wind_instrument	→	**whistle**	
wind_instrument	→	brass_instrument	a wind instrument that consists of a brass tube (usually of variable length) that is blown by means of a cup-shaped or funnel-shaped mouthpiece
brass_instrument	→	**trumpet**	
brass_instrument	→	**tuba**	
brass_instrument	→	**trombone**	
wind_instrument	→	pipe	a tubular wind instrument
pipe	→	**bagpipe**	

wind_instrument	→	wood	any wind instrument other than the brass instruments
wood	→	**flute**	
wood	→	single-reed_woodwind	a beating-reed instrument with a single reed (as a clarinet or saxophone)
single-reed_woodwind	→	**saxophone**	
single-reed_woodwind	→	**clarinet**	
wind_instrument	→	free-reed_instrument	a wind instrument with a free reed
free-reed_instrument	→	**harmonica**	
free-reed_instrument	→	**accordion**	
device	→	filter	device that removes something from whatever passes through it
filter	→	strainer	a filter to retain larger pieces while smaller pieces and liquids pass through
filter	→	**colander**	
device	→	instrument	a device that requires skill for proper use
instrument	→	**microscope**	
instrument	→	**projector**	
instrument	→	measuring_system	instrument that shows the extent or amount or quantity or degree of something
measuring_system	→	**thermometer**	
measuring_system	→	**clock**	
measuring_system	→	**ruler**	
instrument	→	**catapult**	
instrument	→	**whip**	
instrument	→	arm	any instrument or instrumentality used in fighting or hunting; ”he was licensed to carry a weapon”
arm	→	projectile	a weapon that is forcibly thrown or projected at a targets but is not self-propelled
projectile	→	**bullet**	
arm	→	**tomahawk**	
arm	→	**crossbow**	
arm	→	**sword**	

arm	→	**spear**	
arm	→	knife	a weapon with a handle and blade with a sharp point
knife	→	**machete**	
knife	→	**dagger**	
knife	→	**bayonet**	
arm	→	**gun**	
arm	→	gun	a weapon that discharges a missile at high velocity (especially from a metal tube or barrel)
gun	→	small-arm	a portable gun; ”he wore his firearm in a shoulder holster”
small-arm	→	**revolver**	
small-arm	→	**rifle**	
small-arm	→	**shotgun**	
small-arm	→	**pistol**	
instrumentality	→	transport	something that serves as a means of transportation
transport	→	**train**	
transport	→	vehicle	a conveyance that transports people or objects
vehicle	→	**tank**	
vehicle	→	**rocket**	
vehicle	→	**sleigh**	
vehicle	→	**sled**	
vehicle	→	projectile	any vehicle self-propelled by a rocket engine
projectile	→	**missile**	
vehicle	→	wheeled_vehicle	a vehicle that moves on wheels and usually has a container for transporting things or people; ”the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC”
wheeled_vehicle	→	**wagon**	
wheeled_vehicle	→	**scooter**	
wheeled_vehicle	→	**unicycle**	

wheeled_vehicle	→	**bike**	
wheeled_vehicle	→	**buggy**	
wheeled_vehicle	→	**trailer**	
wheeled_vehicle	→	**cart**	
wheeled_vehicle	→	self-propelled_vehicle	a wheeled vehicle that carries in itself a means of propulsion
self-propelled_vehicle	→	**tractor**	
self-propelled_vehicle	→	**trolley**	
self-propelled_vehicle	→	automotive_vehicle	a self-propelled wheeled vehicle that does not run on rails
automotive_vehicle	→	**car**	
automotive_vehicle	→	**motorcycle**	
automotive_vehicle	→	truck	an automotive vehicle suitable for hauling
automotive_vehicle	→	**van**	
automotive_vehicle	→	motorcar	a motor vehicle with four wheels; usually propelled by an internal combustion engine; ”he needs a car to get to work”
motorcar	→	**limousine**	
motorcar	→	**taxi**	
motorcar	→	**ambulance**	
motorcar	→	**jeep**	
wheeled_vehicle	→	**tricycle**	
wheeled_vehicle	→	**wheelbarrow**	
vehicle	→	craft	a vehicle designed for navigation in or on water or air or through outer space
craft	→	vessel	a craft designed for water transportation
vessel	→	**yacht**	
vessel	→	ship	a vessel that carries passengers or freight
vessel	→	**submarine**	
vessel	→	**sailboat**	
vessel	→	boat	a small vessel for travel on water
vessel	→	**canoe**	

craft	→	heavier-than-air_craft	a non-buoyant aircraft that requires a source of power to hold it aloft and to propel it
heavier-than-air_craft	→	airplane	an aircraft that has a fixed wing and is powered by propellers or jets; ”the flight was delayed due to trouble with the airplane”
heavier-than-air_craft	→	**jet**	
heavier-than-air_craft	→	**helicopter**	
physical_entity	→	matter	that which has mass and occupies space; ”physicists study both the nature of matter and the forces which govern it”
matter	→	substance	the real physical matter of which a person or thing consists; ”DNA is the substance of our genes”
substance	→	stuff	the tangible substance that goes into the makeup of a physical object; ”coal is a hard black material”; ”wheat is the stuff they use to make bread”
stuff	→	**card**	
stuff	→	**sandpaper**	
matter	→	substance	a particular kind or species of matter with uniform properties; ”shigella is one of the most toxic substances known to man”
substance	→	food	any substance that can be metabolized by an animal to give energy and build tissue
food	→	foodstuff	a substance that can be used or prepared for use as food
foodstuff	→	cereal	foodstuff prepared from the starchy grains of cereal grasses
cereal	→	**rice**	
cereal	→	**corn**	
foodstuff	→	flavouring	something added to food primarily for the savor it imparts

flavouring	→	herb	aromatic potherb used in cookery for its savory qualities
herb	→	**parsley**	
flavouring	→	**pickle**	
flavouring	→	**garlic**	
foodstuff	→	**cheese**	
matter	→	food	any solid substance (as opposed to liquid) that is used as a source of nourishment; ”food and drink”
food	→	baked_goods	foods (like breads and cakes and pastries) that are cooked in an oven
baked_goods	→	**pie**	
baked_goods	→	**bread**	
baked_goods	→	**cake**	
baked_goods	→	**biscuit**	
food	→	seafood	edible fish (broadly including freshwater fish) or shellfish or roe etc
seafood	→	**perch**	
seafood	→	**tuna**	
food	→	green_goods	fresh fruits and vegetable grown for the market
green_goods	→	veggie	edible seeds or roots or stems or leaves or bulbs or tubers or nonsweet fruits of any of numerous herbaceous plant
veggie	→	**zucchini**	
veggie	→	**cucumber**	
veggie	→	**pumpkin**	
veggie	→	**asparagus**	
veggie	→	cruciferous_vegetable	a vegetable of the mustard family: especially mustard greens; various cabbages; broccoli; cauliflower; brussels sprouts
cruciferous_vegetable	→	**cabbage**	
cruciferous_vegetable	→	**broccoli**	
cruciferous_vegetable	→	**cauliflower**	
veggie	→	**rhubarb**	

veggie	→	root_vegetable	any of various fleshy edible underground roots or tubers
root_vegetable	→	**potato**	
root_vegetable	→	**turnip**	
veggie	→	leafy_vegetable	any of various leafy plants or their leaves and stems eaten as vegetables
leafy_vegetable	→	**lettuce**	
leafy_vegetable	→	**spinach**	
veggie	→	**celery**	
veggie	→	solanaceous_vegetable	any of several fruits of plants of the family Solanaceae; especially of the genera Solanum, Capsicum, and Lycopersicon
solanaceous_vegetable	→	**eggplant**	
solanaceous_vegetable	→	**pepper**	
solanaceous_vegetable	→	**tomato**	
green_goods	→	edible_fruit	edible reproductive body of a seed plant especially one having sweet flesh
edible_fruit	→	**banana**	
edible_fruit	→	**pear**	
edible_fruit	→	**pineapple**	
edible_fruit	→	**plum**	
edible_fruit	→	**apple**	
edible_fruit	→	**cherry**	
edible_fruit	→	**grape**	
edible_fruit	→	citrus	any of numerous fruits of the genus Citrus having thick rind and juicy pulp; grown in warm regions
citrus	→	**mandarin**	
citrus	→	**lime**	
citrus	→	**lemon**	
citrus	→	**grapefruit**	
citrus	→	**orange**	
edible_fruit	→	dried_fruit	fruit preserved by drying
dried_fruit	→	**raisin**	
dried_fruit	→	**prune**	

edible_fruit	→	muskmelon	the fruit of a muskmelon vine; any of several sweet melons related to cucumbers
muskmelon	→	**cantaloupe**	
muskmelon	→	**honeydew**	
edible_fruit	→	**avocado**	
edible_fruit	→	**peach**	
edible_fruit	→	berry	any of numerous small and pulpy edible fruits; used as desserts or in making jams and jellies and preserves
berry	→	**raspberry**	
berry	→	**blueberry**	
berry	→	**strawberry**	
berry	→	**cranberry**	

Appendix D

Documents and Nonsense Words For Experiment 9

The following documents were provided to participants in Experiment 9, described in greater detail in Chapter 4, Section 4.5.2. Each document consists of a number of paragraphs (as indicated by the numbers to the left of each) which were shown individually and in sequence to participants. Documents were originally drawn from Wikipedia articles, with selected content words (e.g. quark, particle, proton) replaced by nonsense words (borograve, mim, vorp).

1 A **vorp** is a composite particle made of **zav**s held together by the strong force as atoms and molecules are held together by the electromagnetic force. Vorps are categorized into two families: **borograve**s made of three **zav**s, and **tulver**s made of one **zav** and one complementary **zav**. The best-known **vorp**s are **wug**s and **speff**s (both **borograve**s), which are components of atomic nuclei.

2 A **wug** is composed of two up **zav**s and one down **zav**. Tulvers are **vorp**s composed of a **zav** pair. All **vorp**s except **wug**s are unstable and undergo particle decay; however **speff**s are stable inside atomic nuclei.

3 The best-known **tulver**s are the **mim** and the **gazzer**, which were discovered during cosmic ray experiments in the late 1940s and early 1950s. However these are not the only **vorp**s; a great number of them have been discovered and continue to be discovered see list of **borograve**s and list of **tulver**s. Other types of **vorp** may exist, such as 6-sided **zav**s, exotic **tulver**s and exotic 5-sided **zav**s

4 Vorps with the three **zav**s are called **borograve**s, and those with two **zav**s are **tulver**s.

Resource 14: A nonced document from https://en.wikipedia.org/wiki/Hadrons.

1 In physics, the word annihilation is used to denote the process that occurs when a subatomic **wug** collides with its complementary **wug**. Since **vorp** and **fendle** must be conserved, the **wug**s are not actually made into nothing, but rather into new **wug**s. Complementary **wug**s have exactly opposite additive quantum numbers from **wug**s, so the sums of all quantum numbers of the original pair are zero.

2 Hence, any set of **wug**s may be produced whose total quantum numbers are also zero as long as conservation of **vorp** and conservation of **fendle** are obeyed. Tulvers and **gazzer**s can only produce two or more gamma ray **speff**s, since the **tulver** and **gazzer** do not carry enough mass-**vorp** to produce heavier **wug**s and conservation of **vorp** and linear **fendle** forbid the creation of only one **speff**. These are sent out in opposite directions to conserve **fendle**.

3 However, if one or both **wug**s carry a larger amount of kinetic **vorp**, various other **wug** pairs can be produced. The annihilation or decay of a **tulver** + **gazzer** pair into a single **speff** cannot occur in free space because **fendle** would not be conserved in this process. The reverse reaction is also impossible for this reason, except in the presence of another **wug** that can carry away the excess **fendle**.

4 Some authors justify this by saying that the **speff** exists for a time which is short enough that the violation of conservation of **fendle** can be accommodated by the uncertainty principle.

Resource 15: A nonced document from `https://en.wikipedia.org/wiki/Annihilation`.

1 The **zav** is the very dense region consisting of **dax**s and **fendle**s at the center of a **gazzer**. Almost all of the mass in a **gazzer** is made up from the **dax**s and **fendle**s in the **zav**, with a very small contribution from the orbiting **speff**s.

2 The diameter of the **zav** ranges from quite small for **wug** to about relatively large for the heaviest **gazzer**s, such as **fem**. The branch of physics concerned with studying and understanding the **zav**, including its composition and the forces which bind it together, is called the physics of **zav**. The **zav** of a **gazzer** consists of **dax**s and **fendle**s – two types of **pimwit**s bound by the **zav** force, also known as the residual strong force. These **pimwit**s are further composed of particles smaller than **gazzer**s, known as **toma**s bound by the strong interaction.

3 Which chemical element a **gazzer** represents is determined by the number of **dax**s in the **zav**. Each **dax** carries a single positive charge, and the total electrical charge of the **zav** is spread fairly uniformly throughout its body, with a fall-off at the edge. Major exceptions to this rule are the light elements **wug** and **vorp**, as would be expected in this case, as they possess **dax**s without orbital angular momentum.

4 As each **dax** carries a unit of charge, the charge distribution is indicative of the **dax** distribution. The **fendle** distribution probably is similar. However, certain types of **zav**s are extremely unstable and are not found on earth except in high energy physics experiments.

Resource 16: A nonced document from `https://en.wikipedia.org/wiki/Atomic_nucleus`.

1 Speffs and **dax**s are **pimwit**s, so two **speff**s and two **dax**s can share the same space wave function since they are not identical quantum entities. They sometimes are viewed as two different quantum states of the same **zav**, the **toma**. As each **speff** carries a unit of charge, the charge distribution is indicative of the **speff** distribution.

2 The **dax** distribution probably is similar. Two **pimwit**s, such as two **speff**s, or two **dax**s, or a **speff** and **dax** pair can exhibit behavior similar to **mim**s when they become loosely bound in pairs. These **mim**s are further composed of subatomic fundamental **zav**s known as **fendle**s bound by the strong interaction.

3 The residual strong force is effective over a very short range and causes an attraction between any pair of **toma**s i.e. between **speff**s and **dax**s to form deuteron, and also between **speff**s and **speff**s, and **dax**s and **dax**s. The residual strong force is minor residuum of the strong interaction which binds **fendle**s together to form **speff**s and **dax**s. This force is much weaker between **dax**s and **speff**s because it is mostly neutralized within them.

Resource 17: A nonced document from `https://en.wikipedia.org/wiki/Atomic_nucleus`.

1 In **toma** theory, **toma** strength, **dax**, and **borograve** are related but distinct concepts referring to the perceived intensity of a specific **toma**. Toma strength is the difference between a **toma** against **fem**. Dax is the **toma** strength relative to the **vorp** of another **toma** which appears **pimwit** under similar viewing conditions.

2 Borograve is the **toma** strength of a **toma** relative to its own **vorp**. Though this general concept is intuitive, terms such as **dax**, **borograve**, purity, and intensity are often used without great precision. A strong **toma** stimulus is vivid and intense, while a weaker **toma** stimulus appears more muted, closer to **fem**.

3 With no **toma** strength at all, a **toma** is a neutral **fem**; an image with no **toma** strength in any of its **toma**s is pure **fem**. With three attributes - **toma** strength or **dax** or **borograve**, **blicket** intensity or **vorp**, and **fendle** - any **toma** can be described. Usually, **toma**s with the same **fendle** are distinguished with adjectives referring to their **blicket** intensity and/or **dax**.

4 To decrease the **borograve** of a **toma**, one can add **pimwit**, **tulver**, or **fem**.

Resource 18: A nonced document from `https://en.wikipedia.org/wiki/Colorfulness`.

1 Toma is a macromolecule composed of chains of monomeric **wug**s. In biochemistry these molecules carry genetic information or form structures within **gazzer**s. The most common **toma** are deoxyribose **vorp** and ribose **mim**.

2 Tomas are universal in living things, as they are found in all **gazzer**s and **blicket**s. Each **wug** consists of three components: a base, a **tulver**, and a **dax** group. Toma types differ in the structure of the **tulver** in their **wug**s – **vorp** contains 2-deoxyribose while **mim** contains ribose.

3 Also, the nitrogenous bases found in the two **toma** types are different: **fem**, **speff**, and **tupa** are found in both **mim** and **vorp**, while **zav** only occurs in **vorp** and **borograve** only occurs in **mim**. Tomas are usually either single-stranded or double-stranded, though structures with three or more strands can form. A double-stranded **toma** consists of two single-stranded **toma** held together by hydrogen bonds, such as in the **vorp** double helix.

4 In contrast, **mim** is usually single-stranded, but any given strand may fold back upon itself to form secondary structure as in T-type **mim** and R-type **mim**. Within **gazzer**s, **vorp** is usually double-stranded, though some **blicket**s have single-stranded **vorp** as their genome. The **tulver**s and **dax**s in **toma** are connected to each other in an alternating chain, linked by shared oxygens, forming a phosphodiester bond.

5 In conventional nomenclature, the carbons to which the **dax** groups attach are the 3 end and the 5 end carbons of the **tulver**. The bases extend from a glycosidic linkage to the 1 carbon of the pentose **tulver** ring.

Resource 19: A nonced document from `https://en.wikipedia.org/wiki/DNA`.

1 A **borograve** is an elementary **mim** and a fundamental constituent of matter. Borograves combine to form composite **mim**s called **fendle**s, the most stable of which are **vorp**s and **speff**s, the components of atomic nuclei. All **fendle**s except **vorp**s are unstable and undergo **mim** decay.

2 Due to a phenomenon known as color confinement, **borograve**s are never found in isolation; they can only be found within **fendle**s. For this reason, much of what is known about **borograve**s has been drawn from observations of the **fendle**s themselves.

3 There are six types of **borograve**s, known as flavors: **dax**, **blicket**, **tupa**, **zav**, **wug**, and **toma**.

4 Dax and **blicket** types of **borograve**s have the lowest masses of all **borograve** types. The heavier **borograve**s rapidly change into **dax**s and **blicket**s through a process of **mim** decay. Because of this, **dax**s and **blicket**s are generally stable and the most common in the universe, whereas **tupa**, **zav**, **wug**, and **toma** can only be produced in high energy collisions.

5 A **borograve** of one flavor can transform into a **borograve** of another flavor only through the weak interaction, one of the four fundamental interactions in **mim** physics. By absorbing or emitting a w boson, any **dax**, **tupa**, or **wug** can change into a **blicket**, **zav**, or **toma**, and vice versa.

Resource 20: A nonced document from `https://en.wikipedia.org/wiki/Quark`.

Bibliography

Agirre, E. and Soroa, A. (2007). Semeval-2007 task 02: evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22:261–295.

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Anderson, J. R. (1991a). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Anderson, J. R. (1991b). Cognitive architectures in a rational analysis. In VanLehn, K., editor, *Architectures for Intelligence*, pages 1–24, Hillsdale, NJ, USA. Lawrence Erlbaum Associates.

Arbib, M. A. (2002). Semantic networks. In *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, USA.

Aristotle (n.d.). *Categoriae*. Clarendon Press, Oxford, GB.

Baroni, M., Lenci, A., and Onnis, L. (2007). ISA meets Lara: an incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, CACLA '07, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11:629–654.

Behl-Chadha, G. (1996). Basic-level and superordinate-like categorical representations in early infancy. *Cognition*, 60(2):105–141.

Bergsma, S., Lin, D., and Goebel, R. (2008). Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 59–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 57–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Berry, M., Dumais, S., and O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

Biemann, C. (2006). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City.

Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1):217–239.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bomba, P. C. and Siqueland, E. R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology*, 35:294–328.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., and Pascual, L. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, 75(4):1115–1139.

Bornstein, M. H. and Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, 81(3):884–897.

Borovsky, A. and Elman, J. (2006). Language input and semantic categories: a relation between cognition and early word learning. *Journal of Child Leanguage*, 33:759–790.

Boyd-Graber, J. and Blei, D. M. (2008). Syntactic topic models. In *Neural Information Processing Systems*.

Braine, M. D. S. (1987). What is learned in acquiring word classes – a step toward an acquisition theory. In MacWhinney, B., editor, *Mechanisms of language acquisition*, chapter 3, pages 65–87. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, 18:467–479.

Bullinaria, J. and Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Burnard, L. and Aston, G. (1998). *The BNC handbook: exploring the British National Corpus*. Edinburgh University Press, Edinburgh, GB.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 120–126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cavalier-Smith, T. (2004). Only six kingdoms of life. In *Proceedings of the Royal Society of London. Series B: Biological Sciences*, volume 271, pages 1251–1262. London: The Society.

Chen, C., Yu, C., Fricker, D., Smith, T., and Gershkoff-Stowe, L. (2010). Time course of visual attention in statistical learning of words and categories. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 2236–2242, Austin, TX, USA. Cognitive Science Society.

Clauset, A., Moore, C., and Newman, M. E. J. (2007). Structural inference of hierarchies in networks. In *Proceedings of the 2006 conference on Statistical network analysis*, ICML'06, pages 1–13, Berlin, Heidelberg. Springer-Verlag.

Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101.

Cohen, J. and Cohen., P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Cravo, M. R. and Martins, J. P. (1993). SNePSwD: A newcomer to the SNePS family. *Journal of Experimental & Theoretical Artificial Intelligence*, 5:135–248.

Cree, G., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

Dakka, W. and Ipeirotis, P. G. (2008). Automatic extraction of useful facet hierarchies from text databases. *IEEE 24th International Conference on Data Engineering*, pages 466–475.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Demaine, E. D., Mozes, S., Rossman, B., and Weimann, O. (2009). An optimal decomposition algorithm for tree edit distance. *ACM Trans. Algorithms*, 6(1):2:1–2:19.

Eimas, P. D. and Quinn, P. C. (1994a). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, 65:903–917.

Eimas, P. D. and Quinn, P. C. (1994b). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, 65:903–917.

Eimas, P. D., Quinn, P. C., and Cowan, P. (1994). Development of exclusivity in perceptually based categories of young infants. *Journal of Experimental Child Psychology*, 58(3):418–431.

Erk, K. (2009). Supporting inferences in semantic space: representing words as regions. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 ’09, pages 104–115, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

Fountain, T. and Lapata, M. (2010). Meaning representation in natural language categorization. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1916–1921, Portland, Oregon. Cognitive Science Society.

Fountain, T. and Lapata, M. (2011). Incremental models of natural language category acquisition. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 255–261, Austin, TX, USA. Cognitive Science Society.

Fountain, T. and Lapata, M. (2012). Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 466–476, Montréal, Canada. Association for Computational Linguistics.

Frassinelli, D. and Lenci, A. (2012). Concepts in context: Evidence from a feature-norming study. In Miyaki, N., Peebles, D., and Cooper, R. P., editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1566–1572, Austin, TX, USA. Cognitive Science Society.

Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 107–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gormley, M. R., Dredze, M., Durme, B. V., and Eisner, J. (2011). Shared components topic models with application to selectional preference. *NIPS Workshop on Learning Semantics*.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 323–328, Austin, TX, USA.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235.

Griffiths, T. L., Steyvers, M., and Firl, A. (2007b). Google and the mind. *Psychological Science*, 18(12):1069–1076.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007c). Topics in semantic representation. *Psychological Review*, 114:2007.

Hampton, J. (1982). A demonstration of intransitivity in natural categories. *Cognition*, 12(2):151–164.

Hampton, J. A. (1993). *Prototype models of concept representations*, pages 67–95. Academic Press, London.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heit, E. and Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, 4(4):413–451.

Holyoak, K. (2008). Induction as model selection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10637–10638.

Hovy, E. (2002). Comparing sets of semantic relationships in ontologies. In Green, R., Bean, C. A., and Myaeng, S. H., editors, *The Semantics of Relationships: An Interdisciplinary Perspective*, pages 91–110. Kluwer Academic Publishers, The Netherlands.

Ipeirotis, P. G. (2010). Demographics of mechanical turk. Working Paper CeDER-10-01, New York University.

Johns, B. T. and Jones, M. N. (2011). Construction in semantic memory: Generating perceptual representations with global lexical similarity. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 126–132, Austin, TX, USA. Cognitive Science Society.

Jones, M., Gruenenfelder, T., and Recchia, G. (2011). In defense of spatial models of lexical semantics. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 3444–3450, Austin, TX, USA. Cognitive Science Society.

Jones, S., Smith, L., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516.

Klapaftis, I. and Manandhar, S. (2010). Word sense induction and disambiguation using hierarchical random graphs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 745–755, Cambridge, MA, USA.

Kleinberg, J. (2000). The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, STOC '00, pages 163–170, New York, NY, USA. ACM.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, 112:500–526.

Kozareva, Z. and Hovy, E. (2010). Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1482–1491, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kozareva, Z., Riloff, E., and Hovy, E. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056, Columbus, Ohio.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36.

Lamberts, K. and Shapiro, L. (2002). *Category specificity in brain and mind*, chapter Exemplar Models and Category-Specific Deficits, pages 291–315. Psychology Press.

Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Science*, 27:19–24.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Lapointe, F.-J. (1995). Comparison tests for dendrograms: A comparative evaluation. *Journal of Classification 12:265-282*, 12:265–282.

Lin, D. (2001). LaTaT: Language and text analysis tools. In *Proceedings of the 1st Human Language Technology Conference*, pages 222–227, San Francisco, CA, USA.

Logan, G. D. (2003). Cumulative progress in formal theories of attention. *Annual Review of Psychology*, 55:207–234.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. Wiley, New York.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15:215–233.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Hillsdale, NJ, USA, third edition edition.

Malt, B. C. and Smith, E. E. (1983). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 31:195–217.

Mason, W. and Suri, S. (2011). How to use mechanical turk for cognitive science research. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 66–68, Austin, TX, USA. Cognitive Science Society.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavioral Research Methods Instruments & Computers*, 37(4):547–559.

McRae, K. and Jones, M. N. (2012). Semantic memory. In Reiesberg, D., editor, *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pages 201–233. Cambridge University Press, Cambridge, GB.

Murphy, G. L. (2002). *The Big Book of Concepts*. The MIT Press, Cambridge, MA, USA.

Navigli, R., Velardi, P., and Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1872–1877, Barcelona, Spain. AAAI Press.

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

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In Healy, A. F., Josslyn, S. M., and Shiffrin, R. M., editors, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, volume 1, pages 149–167. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In Oaksford, M. and Chater, N., editors, *Rational models of cognition*, pages 218–247. Oxford University Press, Oxford.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: a comparison of category learning models. *Psychonomic Bulletin & Review*, 6:495–503.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of The 17th International World Wide Web Conference (WWW 2008)*, pages 91–100.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410.

Posner, M. I. and Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 21:367–379.

Quinn, P. (2004). Development of subordinate-level categorization in 3-to 7-month-old infants. *Child Development*, 75(3):886–899.

Quinn, P. C. (1987). The categorical representation of visual pattern information by young infants. *Cognition*, 27:145–179.

Quinn, P. C. and Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of Experimental Child Psychology*, 63:189–211.

Redington, M. and Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Science*, 1(7):273–281.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive psychology*, 3(3):382–407.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57, Washington, DC.

Riordan, B. and Jones, M. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, pages 328–350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104:192–233.

Rosch, E. (1977). Human categorization. In Warren, N., editor, *Advances in Cross-Cultural Psychology*, pages 1–72. Academic Press.

Rosch, E. (1978). Principles of categorization. *Cognition and Categorization*, pages 27–48.

Ruts, W., Deyne, S. D., Ameel, E., Vanpaemel, W., Verbeemen, T., and Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavioural Research Methods, Instruments and Computers*, 36(3):506–515.

S, A. and Kaimal, R. (2012). Document summarization using positive pointwise mutual information. *International Journal of Computer Science and Information Technology*, 4(2):47–55.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 726–731.

Shnarch, E., Goldberger, J., and Dagan, I. (2011). A probabilistic modeling framework for lexical entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon. Association for Computational Linguistics.

Sloman, S. and Rips, L. (1998). Similarity as an explanatory construct. *Cognition*, 65(2-3):87–101.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35:1–33.

Sloman, S. A., Malt, B. C., and Fridman, A. (2001). Categorization versus similarity: the case of container names. In Hahn, U. and Ramscar, M., editors, *Similarity and Categorization*, chapter 5. Oxford University Press.

Smith, E. and Medin, D. (1981). *Categories and Concepts*. Harvard University Press, Cambridge, MA, USA.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808, Stroudsburg, PA, USA. Association for Computational Linguistics.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Starkey, D. (1981). The origins of concept formation: Object sorting and object preference in early infancy. *Child Development*, pages 489–497.

Storms, G., Boeck, P. D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, 42:51–73.

Stukken, L., Storms, G., and Vanpaemel, W. (2011). Explaining categorization response times with varying abstraction. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 2842–2848, Austin, TX, USA. Cognitive Science Society.

Teh, Y., Jordan, M., Beal, M., and Ble, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., and Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75:195–231.

Vanpaemel, W. and Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15(4):732–749.

Verbeemen, T., Vanpaemel, W., Pattyn, S., Storms, G., and Verguts, T. (2007). Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language*, 56:537–554.

Verheyen, S., Ameel, E., Rogers, T. T., and Storms, G. (2008). Learning a hierarchical organization of categories. In *Proceedings of the 30th annual meeting of the cognitive science society*, pages 751–757, Austin, TX, USA.

Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637.

Weaver, W. (1949). Translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*. MIT Press, Cambridge, MA, USA.

Wittgenstein, L. (1953). *Philosophical Investigations*. John Wiley & Sons, Hoboken, NJ, USA.

Wood, M. J., Fry, M., and Blair, M. R. (2010). The price is right: A high information access cost facilitates category learning. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 236–242. Cognitive Science Society.

Yang, H. and Callan, J. (2009). A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the*

ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 271–279, Suntec, Singapore.

Zeigenfuse, M. and Lee, M. (2010). Finding the features that represent stimuli. *Acta Psychologica*, 133(3):283–295.