598

ENCYCLOPEDIA OF COGNITIVE SCIENCE

2000

Prototype representations

James A. Hampton

City University London UK

Contents list

Prototype representations represent classes in terms of their central or most typical example – the prototype – rather than in terms of an explicit definition of the class boundary. Classification is based on whether similarity to the prototype is above some threshold criterion value.

Prototype Representations

Introduction

Psychological models of conceptual knowledge and knowledge of word meaning both require an internal representation of the world in the mind. Models or systems of representation vary in their level of complexity. At the simplest level are exemplar storage systems, in which individual experiences of objects are each stored, and categorization of novel instances proceeds by comparison to previously encountered exemplars. At the most complex level are systems for representing complex knowledge incorporating causal explanations of how individual attributes of an object class relate to each other. Prototype representations lie in the middle of this complexity dimension. The central notion is that we abstract a generic representation or prototype of a class (such as Fruit) from our experience with many examples. This representation may be more or less structured, but does not contain information about specific individuals. Deciding that a novel instance is of a particular type involves a decision about how closely it matches the prototype for the class. This form of representation gives rise to two phenomena. First the membership of the class is graded in terms of typicality. Those instances that match the prototype well (such as an apple) will be considered more representative or typical than those that match poorly (such as a coconut). Second, because the borderline of the class is not well specified, there may be instances which are neither clearly in, nor clearly out of the class (such as tomatoes or olives). This vagueness in the application of the category is a key aspect of prototype representations since it maps onto the acknowledged vagueness of the meaning of most nouns and verbs in natural language.

Historical Developments

Early theories of semantic and conceptual representation tended to assume that a conceptual category was defined in terms of a set of singly necessary and jointly sufficient defining features – a view known as the Classical View of concept structure. It can be seen in analyses of the meaning of words like “Uncle” or “Cousin”, which can be defined in terms of a small set of dimensions.

One of the first appearances of the notion of prototype was a classic study by Posner & Keele (1968). Their study investigated the learning of novel stimulus classes through trial and error learning. Rather than investigating the learning of different forms of definitional rule, Posner and Keele taught people to differentiate stimulus classes that were based on small distortions of a random array of dots. In general terms a prototype can be thought of as a point in "stimulus space". Each exemplar of a category can be described in terms of its position along a number of orthogonal dimensions (for example its size, orientation, colour etc). By plotting the dimensions in an multi-dimensional space, the exemplars can be placed in the space according to their co-ordinates on each dimension, and distance in the space can then be mapped onto the similarity between exemplars. This spatial representation is the stimulus space. The prototype is then defined as the centre of gravity of the set of exemplars in the space – the (possible) exemplar that has maximum overall similarity to all of the surrounding exemplars.

(Where a stimulus is not describable in terms of continuous or dichotomous dimensions, then this spatial model breaks down, and other ways of defining the prototype are used, such as taking the prototype to be the possible exemplar that has the most commonly occurring values on each feature or dimension.)

Prototype representations have in common that they are limited to representing classes that are linearly discriminable. That is to say that it is possible to discriminate between members and non-members in terms of some simple additive combination of the features or dimensions. Probability of an item being categorized in a prototype class should be expressible as a positive function of its similarity to the prototype and as a negative function of its similarity to the prototypes of other contrasting categories. This constraint has been used to argue that natural biological categories are not represented by prototypes. For example, it is claimed, a creature such as a whale is more similar to the prototype for fish than is a creature such as a seahorse, and yet while the whale is not a fish, the seahorse is.

The primary development of prototype theory is largely due to the work of Eleanor Rosch and Carolyn Mervis in the 1970s. Rosch (working initially as Heider) first developed the notion of "natural" prototypes in the context of colour and geometric figures. She showed that in the colour spectrum there were certain hues that were naturally considered "good examples" of a colour term such as "red", and others that were atypical. In her work with a primitive society – the Dani of New Guinea – Rosch was able to show that even although the group had no language term for "red", they found colour categories based on a "good" red easier to learn than colour categories centred around a "poor" or atypical red. (More recent research has questioned this result). According to Rosch, certain visual forms and colours form natural (universal) cognitive reference points or prototypes. Work in comparative linguistics confirms this notion. Whereas the boundaries between colour terms vary widely across languages, the choice of the most central hue for a colour term shows much closer agreement.

Family resemblance structures: Wittgenstein and Rosch

In subsequent work with Mervis, Rosch extended the notion of prototype concept to include more complex natural categories such as biological and artefact kinds (birds, fish, fruits, tools, furniture etc.) In developing the notion, they referred to a philosophical analysis given by Wittgenstein, 1953 in his *Philosophical Investigations*. Wittgenstein spent many years worrying over the relation between word meaning and the underlying logic of thought and language. Towards the end of his life he came to the view that words do not correspond directly to the logical terms of propositions, but that the meaning of a word is defined instead in terms of the complex pattern of use that it has in language. He pointed out, for example, that one cannot specify just what all the activities that we call "game" have in common, since for any possible relevant defining feature (e.g. "games are all competitive") clear counterexamples could be found (ring-a-rosy is a children's game that is not competitive). Games, he said, are like members of a family. There are clear resemblances amongst the members of a family, but there may be no single distinguishing feature that would pick out the whole family from everyone else.

Rosch & Mervis (1975) took this notion and developed it into the Prototype Theory of Concepts (PTC). According to PTC, a concept such as game is defined in terms of a number

of features, such as *competitive* or *has teams*, and membership in the class of games involves possessing a sufficient number of these features. Their theory was equivalent to a classification system known as polythetic classification in theoretical taxonomy. The theory proposes that because different features or dimensions of objects are correlated within a domain, objects naturally fall into similarity clusters. For example in the domain of creatures having a beak is correlated with laying eggs and flying, while having a scaly skin is correlated with having cold blood and teeth. Prototype representations capitalise on these inter-correlations amongst features by drawing clusters of inter-correlated features together into a prototype of, say, birds or reptiles. However because the correlations are imperfect, knowing that an object is in a particular class would not tell you just which set of the prototypical features it would possess.

Evidence for their theory was obtained as follows. They presented subjects with members of categories such as Birds or Fruits, and had them list properties or features of those members. Other subjects then decided the extent to which each member possessed each feature. A tally was computed of the degree to which each member shared features with each other member. This "family resemblance score" was shown to correlate well with independent judgements of the typicality of each member. Using a more direct method of interrogating subjects about the features that they thought relevant to defining each category, Hampton confirmed that the number of shared features predicts typicality. He also showed that it could be used with some degree of accuracy to determine whether an item was considered to be a category member or not, and the speed with which the decision was made.

The novel feature of PTC as applied to word meanings is that, unlike traditional lexical semantic analyses of meaning, a feature may be part of the "definition" of a term even though it is not true of all the things covered by the term. Analyses of prototype effects in any domain typically involve showing four types of effect:

- There is no explicit definition to be given in terms of a conjunction of defining features
- Features (e.g. that birds can fly) are listed as important to the concept's meaning even though they are not in fact common to all members of the category
- There are clear differences in the "representativeness" or typicality of different category members
- There are differences in the degree to which different items may actually be considered to belong in the category – that is to say the category borderline is vague or fuzzy.

Rosch, Mervis, and others extended the exploration of prototype effects into many different areas. Prototype effects have been found in the learning of novel categories based on prototypes, the speed and accuracy of categorization, the strength of inductive inferences based on a typical as opposed to an atypical category member, and the differential build up and release from proactive interference in short term memory studies. An example of the predictive power of PTC is a study of categorization by (Hampton, 1982). One consequence of defining category membership in terms of a "sufficient number" of the prototype features is that one may find a hierarchy of classes (A is a kind of B; B is a kind of C) for which the transitive inference (A is a kind of C) does not hold. Hampton confirmed the existence of such sets. For example "car-seat" was categorized as a chair, and "chair" was categorized as furniture, but "car-seat" was not considered to be furniture.

Another closely related example of the use of prototype representations in reasoning is the conjunction fallacy, in which people erroneously use similarity to a prototype to estimate probabilities of class membership.

PTC as proposed by Rosch is not in itself a psychological model of concept representation, but is more a way of drawing together a large set of phenomena. More precise models for representing and learning prototype concepts have since been developed. While PTC has undoubtedly had a great influence on the understanding of concepts and word meaning, it is important to note that few theorists still regard it as adequate as a basis for conceptual representation. Murphy & Medin (1985) developed a critique of the theory from the point of

view of the lack of constraints that it provides on the notion of feature or similarity. They argue that one cannot define similarity without making prior assumptions about the relevant dimensions of difference amongst a set of stimuli, and that these assumptions come from a much more sophisticated understanding of the domain in question. Concepts have to be seen in the context of a broader domain "theory" in which they play an explanatory role. This view, sometimes known as the "theory theory" has been widely endorsed by developmentalists studying children's concepts (e.g. Carey, Gopnik, Keil).

Importance of prototype theory for investigations into word meaning

Evidence for prototype effects in word use and word meaning is easy to find. In addition to the intransitivity of categorization, the perceived strength of inductive arguments is also affected by typicality differences. Even when told that all members of a broad category have some property, people feel more confident about concluding that a typical subclass would have it, than that an atypical subclass would. People are aware that the more typical an example, the more likely it is to possess all the prototypical features of the class.

Demonstrations of the vagueness of category borderlines are particularly important in studies of word meaning. It can be argued that all natural language terms (that is, excluding terms in axiomatically defined systems such as mathematics) are vague to some degree. Judgements on whether to call a man tall, whether to call a geological formation a mountain, or whether to call a particular organism a dog may all be a matter of debate. There is some evidence however that with biological kinds as opposed to artefacts or social categories, people are less willing to agree that the boundaries of the class may be vague, and instead assume that there is some essential constitutive property, known perhaps only to experts. People for example may assume that "arthritis" refers to some well-defined condition with a unique identifiable cause, whereas in fact it simply means inflammation of the joints.

A simple way to account for the ubiquitous vagueness of natural language terms is through Wittgenstein's basic insight that, as Rosch puts it, we can "judge how clear a case something is and deal with categories on the basis of clear cases in the total absence of information about boundaries". Prototype representations can be held independently of a rule for determining the category boundary.

Take an example such as "murder". Most people's understanding of the meaning of this word in English is based on prototypical examples – the classical murder mystery type of murder in which one individual deliberately and intentionally kills another through their own direct physical action, with some clear motive such as revenge, jealousy or personal financial gain. A prototype analysis of the concept would involve gathering a list of all such features from a sample of English speakers and putting them together to generate a prototypical or paradigm case. From this prototype one can then invent different possible scenarios in which one or more of the standard features are missing. These might include cases where the killing occurs through failing to act, cases where the victim may be considered a borderline case of being a person (as in abortion) or cases where the motives are not self-serving (as in euthanasia). The doubt and debate that is engendered in this series of "moral dilemmas" are evidence of the multi-dimensional nature of the concept itself and the vagueness that results from representing clear cases, but not clearly representing the class boundary.

The problems of vagueness in language use provide a good argument for a clear differentiation between concepts and word meanings. Osherson & Smith (1981) and a number of philosophers have argued that our ability to understand and to use logic in our thinking and speaking is itself evidence that concepts cannot be vague prototypes. Osherson and Smith, for example, argue that there is no workable logic for handling logical combinations of vague concept terms such as "striped apple" or "pet fish". Typicality in these complex concepts is not a simple function of typicality in the original sets, since a guppy or goldfish may be a clear example of a pet fish, but atypical as either a pet or a fish. Work on conceptual combination arising from their paper has shown that the combination of prototypes does obey some constraints, although there is also evidence that broader world knowledge is involved in determining the meaning of complex noun phrases. The difficulty is that if concepts are to be components of thoughts, then the way in which they combine should follow the simple rules of logic. Since prototypes are mostly non-compositional, it is argued

that they cannot serve the necessary function of being the building blocks for a compositional theory of thought.

Another important critique of PTC was offered by Armstrong, Gleitman, and Gleitman (1983), who asked subjects to give typicality ratings to exemplars of well-defined categories such as "even number". The degree of inter-subject agreement on the typicality of numbers such as 2, 18 or 574 was as great as that for the typicality of different fruits or different items of furniture. From this result they argued that typicality effects per se were not strong enough evidence to support the conclusion that a concept had a prototype representation. At the very least, their demonstration makes the point that determining the membership of a class need not involve the same information as judging what is typical of that class.

Prototype theory has been applied to the analysis of a wide range of semantic domains. Labov demonstrated how the use of simple terms such as "cup" and "bowl" could be mapped into a stimulus space involving dimensions of size, shape and use of containers. Cantor and colleagues have applied the analysis with success to person perception, personality traits, psychological situations and psychiatric diagnosis. Coleman and Kay showed graded structure in the types of speech act that would be categorized as "lying", and Hampton applied it to abstract terms such as Art and Science, although interestingly some other terms like Rule and Instinct did not show prototype structure.

Prototypes in Linguistics, and radial categories in word meaning

Cognitive linguistics has embraced the notion of prototypes for analysis of word usage and word meaning. Most notable has been the work of George Lakoff, whose book *Women. Fire, and Dangerous Things* ((Lakoff, 1987)) presents a detailed account of prototype effects, which are viewed as reflecting the underlying idealised cognitive models that we use to represent the world. According to Lakoff, prototype effects can be identified not only in lexical meaning, but also in phonology, morphology and syntax. For example Ross has argued for the prototype nature of syntactic categories such as noun. Some nouns (typically referring to concrete objects) show a wider range of allowable syntactic manipulations than do others. Membership in the category appears to be graded – some words are more "nouny" than others.

Lakoff distinguishes a number of different sources of prototype effects in lexical semantics. One case is where a number of related cognitive models cluster within the same domain. For example the concept "mother" really involves a cluster of concepts including giving birth, being the genetic parent, nurturing the child, and playing the relevant family role. Surrogate, adoptive, foster and egg-donor mothers fit one or another of these models. Prototype effects result from the different applicability of the cluster of models to different cases.

Lakoff takes the analysis further with his discussion of "radial categories". These are clusters of meaning in which the application of a term has come to be extended to a range of other cases through a non-arbitrary but yet unpredictable process of chaining. For example in Japanese, there is a noun classifier "hon" which is used most commonly to describe long thin objects. By radial chaining it is also used for associated nouns such as martial arts contests using staffs or swords, hits in baseball, telephone calls through long thin wires, injections using long thin needles and so forth. According to Lakoff the best explanation for the diverse range of nouns that take "hon" is in terms of chains of association in which a central case (e.g. a long thin staff) becomes extended to a secondary case (e.g. a martial arts contest) and from there to a case at third remove such as a judo contest (similar to a martial arts contest, but now lacking the long thin staff).

As another example of prototype effects in language, Lakoff also reports an extensive analysis by Brugman of the different uses in English of the spatial preposition "over". Although at first sight a word with only one meaning, a careful analysis reveals a host of different but related senses – moving above and across (jumping over), being above (hovering over), covering up (painting over), to name but three. It is argued that the way in which word meanings form an interconnected network of related senses shows a particular form of prototype structure that is endemic to natural languages. The process of chaining has also been shown to affect categorization in the learning of novel categories.

One consequence of radial category structure is that the original notion of representing a concept with a single prototype and an allowable degree of distortion is no longer adequate. Because the radial extensions take the meaning along chained paths in unpredictable ways, one can no longer expect the application of a term to occupy a linearly discriminable region of the stimulus space. A good demonstration of the complexity of radial categories is Malt's analysis of the use of the word "water". Malt asked a group of students to judge the extent to which a range of liquids such as lemonade, dishwater, rain etc. contained H_2O. There was surprisingly little correlation between the estimated proportion of H_2O and the appropriateness of the term "water" for referring to the liquid. Use of the term was affected by a number of other factors such as how the liquid was used, and whether other more appropriate labels for it existed.

Acquisition of prototype categories

The acquisition of prototype categories has been studied in two different ways. First there have been extensive studies of how adults (and children) learn novel categories based around a prototype structure. The domains used include a wide range of materials such as stick figures, random shapes, simple geometric figures, lists of disease symptoms, random strings of letters or schematic scenes. Prototype models make the specific prediction that even if not presented in training, the prototype stimulus itself will always be at least as fast and as accurately classified as other category members. The main competitor to prototype theory for category learning is the Generalised Context Model (GCM) which assumes that individual exemplars are stored without any abstraction of the prototype. Across a range of experiments, the GCM has been developed into a highly successful predictor of a range of results, and frequently out-performs prototype models in predicting behaviour. However the two approaches are perhaps best seen as variants of a more general model. Both learn the statistical properties of the stimulus input, in a way that could easily be modelled by a neural net with a hidden layer of nodes between the input of features and the output of category membership. As the hidden layer becomes more restricted, so the ability to retain individual exemplar information is lost and abstraction of a prototype becomes more important.

The second way in which the acquisition of prototypes has been studied is through the study of children's early use of words. Keil & Batterman (1984) presented children with a range of concepts such as island or uncle, and tested how they would categorize novel examples. Younger children aged 4-5 years tended to categorize on the basis of surface appearance – for example an "uncle" is an adult male who gives you presents on your birthday. Older children changed to using more definitional information – for example allowing that if your grandmother had a young male child then that could still be your uncle. Keil accounted for this change by suggesting that children start out forming similarity based prototype concepts for the meaning of these terms based on the examples that they have experienced. It is only later, as they start to develop causal explanatory principles for organising their knowledge, that they switch to the correct adult usage of the terms.

Cross linguistic analyses

Schwanenflugel, Blount, and Lin (1991) review the influence of cultural and linguistic differences on word meanings and concepts. The culture and language into which a child is born presents her with a system of cutting up the world and labelling it which requires attention to the correct attributes and dimensions, and learning of the appropriate underlying theories and models of different domains.

Studies of ethnobiological terms (words for biological kinds) suggest that all cultures divide up the natural world in similar ways, and have labels corresponding roughly to the level of species. In other domains there are however important cross-cultural differences in how language divides up the world. For example, Polish has no word corresponding to the English word "disgust", and English no word corresponding to the German "Gemutlichkeit". Although sometimes taken as evidence that thought is thereby constrained by language, the fact that we frequently adopt useful terms from foreign languages into our own vocabulary suggests that the constraint can be overcome. A language that lacks a term for a particular concept is very similar to an individual language speaker whose vocabulary in her native

language is limited. There is always the possibility of extending the expressive range of one's language.

A study by Malt and Sloman investigated the domain of container names in English, Spanish and Chinese Mandarin. In one task, participants were given a set of photographs of a wide variety of different containers (boxes, cartons, bottles, jars etc.) and had to say what they would call them. The results showed almost no relationship between one language and another. A word in one language would often pick out a class that cut across several classes in another language. However when another group of participants rated the similarity of each container to the others, there was far greater agreement between the different language groups. It is therefore clear that the effects of language and the effects of the cultural environment in which a person lives may be independent of each other. Rated similarity was based on a shared experience of the appearance and use of a particular container, whereas the label given to it was highly idiosyncratic to the particular language being spoken.

Another case in which languages differ markedly is in the use of spatial prepositions such as "in", "over" or "through" in English, which according to Lakoff form radial categories of meaning. Even European languages with common roots such as French Italian and Spanish have very different ways in which these terms map onto the world, and learning their use requires attention to subtly different conceptual distinctions in each language.

Reference List

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

Hampton,J.A. (1982). A Demonstration of Intransitivity in Natural Categories. *Cognition, 12*, 151-164.

Keil,F.C., & Batterman,N. (1984). A Characteristic-to-Defining Shift in the Development of Word Meaning. *Journal of Verbal Learning and Verbal Behavior, 23*, 221-236.

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

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

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

Posner,M.I., & Keele,S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

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

Schwanenflugel,P., Blount,B.G., & Lin,P. (1991). Cross-Cultural Aspects of Word Meanings. In P. Schwanenflugel (Ed.), *The Psychology of Word Meanings*. (pp. 71-90). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wittgenstein,L. (1953). *Philosophical Investigations*. New York: Macmillan.

Glossary

Prototype#Most central or representative exemplar of a class

Semantic#To do with the meaning of words

Representation#A mental entity that encodes information about the outside world

Concept#An element of thought, used to categorize and understand the world

Conjunction fallacy#The belief that an individual may be more likely to belong in a conjunction of two classes than in one of the classes alone.

Exemplar#An individual case that belongs in a particular category

Categorization#Classification of individual cases into different categories

Further Reading

Schwanenflugel,P.J. (Ed.). (1991). *The Psychology of Word Meanings*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Mechelen,I., Hampton,J.A., Michalski,R.S., & Theuns,P. (Eds.). (1993). *Categories and Concepts: Theoretical Views and Inductive Data Analysis*. London: Academic Press.

Berlin,B., Breedlove,D.E., & Raven,P.H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist, 75*, 214-242.

Goldstone,R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125-157.

Hampton,J.A. (1995). Testing Prototype Theory of Concepts. *Journal of Memory and Language, 34*, 686-708.

Hampton,J.A. (1997). Psychological representation of concepts. In M. A. Conway (Ed.), *Cognitive Models of Memory*. Hove: Psychology Press.

Kamp,H., & Partee,B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129-191.

Keefe,R., & Smith,P. (1997). Theories of vagueness. In R. Keefe & P. Smith (Eds.), *Vagueness: a reader*. (pp. 1-57). Cambridge: MIT Press.

Rosch,E., & Lloyd,B.B. (1978). *Cognition and Categorization*. Hillsdale NJ: Lawrence Erlbaum Associates.

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