 Vol. 2(2), pp. 9–27

Conceptual Spaces as a Framework for Knowledge Representation

Peter Gärdenfors
Department of Cognitive Science
Lund University, Sweden

Abstract

The dominating models of information processes have been based on symbolic representations of information and knowledge. During the last decades, a variety of non-symbolic models have been proposed as superior. The prime examples of models within the non-symbolic approach are neural networks. However, to a large extent they lack a higher-level theory of representation. In this paper, conceptual spaces are suggested as an appropriate framework for non-symbolic models. Conceptual spaces consist of a number of "quality dimensions" that often are derived from perceptual mechanisms. It will be outlined how conceptual spaces can represent various kind of information and how they can be used to describe concept learning. The connections to prototype theory will also be presented.

1. The Problem of Modeling Representations

Cognitive science has two overarching goals. One is *explanatory*: By studying the cognitive activities of humans and other animals, one formulates *theories* of different aspects of cognition. The theories are tested by experiments or by computer simulations. The other goal is *constructive*: By building *artifacts* like chess-playing programs, robots, animats, etc., one attempts to construct systems that can accomplish various cognitive tasks. For both kinds of goals, a key problem is how the *representations* used by the cognitive system are to be modeled in an appropriate way.

In cognitive science, there are currently two dominating approaches to the problem of modeling representations. The *symbolic* approach starts from the assumption that cognitive systems should be modeled by Turing machines. On this view, cognition is seen as essentially involving symbol manipulation. The second approach is *associationism*, where associations between different kinds of information elements carry the main burden of representation. Connectionism is a special case of associationism, which models associations by artificial neuron networks. Both the symbolic and the associationistic approaches have their advantages and disadvantages. They are often presented as competing paradigms, but since they attack

cognitive problems on different levels, I shall argue later that they should rather be seen as complementary methodologies.

However, there are aspects of cognitive phenomena for which neither symbolic representation nor connectionism seem to offer appropriate modeling tools. In this article, I will advocate a third form of representing information that is based on using *geometrical* structures rather than symbols or connections between neurons. Using these structures *similarity* relations can be modeled in a natural way. The notion of similarity is crucial for the understanding of many cognitive phenomena. I shall call my way of representing information the *conceptual* form since I believe that the essential aspects of concept formation are best described using this kind of representations.

Again, conceptual representations should not be seen as competing with symbolic or associationist (connectivist) representations. Rather, the three kinds can be seen as three *levels* of representations of cognition with different scales of resolution.

I shall outline a theory of *conceptual spaces* as a particular framework for representing information on the conceptual level. A conceptual space is built up from geometrical representations based on a number of *quality dimensions*. The emphasis of the theory will be on the constructive side of cognitive science. However, I believe that it also can explain several aspects of what is known about representations in various biological systems.

2. Quality Dimensions

One notion that is severely downplayed in symbolic representations is that of *similarity*. I submit that judgments of similarity are central for a large number of cognitive processes. Judgments of similarity reveal the dimensions of our perceptions and their structures. For many kinds of dimensions it will be possible to talk about *distances*. The general assumption is that the smaller the distances is between the representations of two objects, the more similar they are. In this way, the similarity of two objects can be defined via the distance between their representing points in the space. Thus conceptual spaces provide us with a natural way of representing similarities. In general, the epistemological role of the conceptual spaces is to serve as a tool in sorting out various *relations* between perceptions.

As introductory examples of quality dimensions one can mention temperature, weight, brightness, pitch and the three ordinary spatial dimensions height, width and depth. I have chosen these examples because they are closely connected to what is produced by our sensory receptors (Schiffman 1982). The spatial dimensions height, width and depth as well

as brightness are perceived by the visual sensory system, pitch by the auditory system, temperature by thermal sensors and weight, finally, by the kinesthetic sensors. There are additional quality dimensions that are of an abstract, non-sensory character.

The primary function of the quality dimensions is to represent various "qualities" of objects.[1] They correspond to the different ways stimuli are judged to be similar or different. In most cases, judgments of similarity and difference generate an ordering relation of stimuli (Clark 1993, p. 114). For example, one can judge tones by their pitch which will generate and ordering of the perceptions. The dimensions form the "framework" used to assign *properties* to objects and to specify *relations* between them. The coordinates of a point within a conceptual space represent particular instances of each dimension, for example a particular temperature, a particular weight, etc.

The quality dimensions are taken to be independent of symbolic representations in the sense that we and other animals can represent the qualities of objects, for example when planning an action, without presuming an internal language or another symbolic system in which these qualities are expressed. In other words, the dimensions are the building blocks of representations on the conceptual level.

When the explanatory aim of cognitive science is in focus, the quality dimensions should be seen as theoretical entities used as a modeling factor in describing cognitive activities of organisms. When constructing artificial systems, the dimensions function as the framework for the representations used by the systems.

The notion of a dimension should be understood literally. It is assumed that each of the quality dimensions is endowed with certain *geometrical* structures (in some cases they are *topological* or *orderings*). As a first example to illustrate such a structure, Fig. 1 shows the dimension of "weight" which is one-dimensional with a zero point, and thus isomorphic to the half-line of non-negative numbers. A basic constraint on this dimension that is commonly made in science is that there are no negative weights.[2]

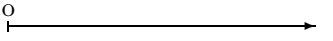

Figure 1: The weight dimension.

[1] In traditional philosophy, following Locke, a distinction between "primary" and "secondary" qualities is often made. This distinction corresponds roughly to the distinction between "scientific" and "phenomenal" dimensions to be presented in the following section.

[2] However, it is interesting to note (cf. Kuhn 1970) that during a period of phlogiston chemistry, scientists were considering negative weights in order to evade some of the anomalies of the theory.

In previous writings on conceptual spaces, I have used the example of the perceptual color space to illustrate a more structured set of quality dimensions (Gärdenfors 1990, 1991, 2000). However, we can also find related spatial structures for other sensory qualities. For example, consider the quality dimension of *pitch*, which is basically a continuous one-dimensional structure going from low to high tones. This representation is directly connected to the neurophysiology of pitch perception.

Apart from the basic frequency dimension of tones, it is possible to identify some further structure in the mental representation of tones. Natural tones are not simple sinusoidal tones of only one frequency, but are constituted of a number of higher harmonics. The timbre of a tone, which is a phenomenal dimension, is determined by the relative strength of the higher harmonics of the fundamental frequency of the tone. An interesting perceptual phenomenon is "the case of the missing fundamental". If the fundamental frequency is removed by artificial methods from a complex physical tone, the phenomenal pitch of the tone is still perceived as that corresponding to the removed fundamental.[3] Apparently, the fundamental frequency is not indispensable for pitch perception, but the perceived pitch is determined by a combination of the lower harmonics.

Thus, the harmonics of a tone are essential for how it is perceived. This entails that tones which share a number of harmonics will be perceived to be similar. The tone that shares the most harmonics with a given tone is its octave, the second most similar is the fifth, the third most similar is the fourth and so on. This additional "geometrical" structure on the pitch dimension, which can be derived from the wave structure of tones, provides the foundational explanation for the perception of musical *intervals*.[4]

For another example of sensory space representations let me only mention that the human perception of *taste* appears to be generated from four distinct types of receptors: saline, sour, sweet, and bitter. Thus the quality space representing tastes could be described as a 4-dimensional space. One such model was put forward by Henning (1961), who suggested that phenomenal gustatory space could be described as a tetrahedron (see Fig. 2). Actually, Henning speculated that any taste could be described as a mixture of only three primaries. This means that any taste can be represented as a point on one of the *planes* of the tetrahedron, so that no taste is mapped onto the interior.

However, there are other models which propose more than four fundamental tastes.[5] The best model of the phenomenal gustatory space remains to be established. This will involve sophisticated psychophysical

[3] See e.g. Gabrielsson (1981), pp. 20–21.

[4] For some further discussion of the structure of musical space see Gärdenfors (1988), Sections 7–9.

[5] See Schiffman (1982), Chap. 9, for an exposition of some such theories.

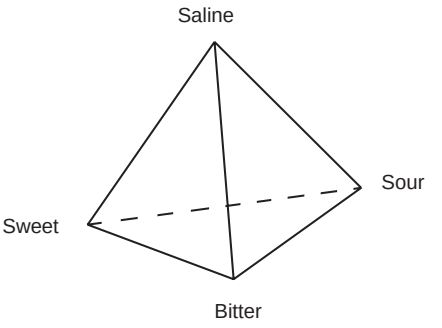

Figure 2. Henning's taste tetrahedron.

measurement techniques. Suffice it to say that the gustatory space quite clearly has some non-trivial geometrical structure. For instance, we can meaningfully claim that the taste of a walnut is *closer* to the taste of a hazelnut than to the taste of popcorn in the same way as we can say that the color orange is closer to yellow than to blue.

It should be noted that some quality "dimensions" have only a *discrete* structure, that is, they merely divide objects into disjoint classes. Two examples are classifications of biological species and kinship relations in a human society. One example of a phylogenetic tree of the kind found in biology is shown in Fig. 3. Here the nodes represents different species in the evolution of, for example, a family of organisms, where nodes higher up in the tree represent evolutionarily older (extinct) species.

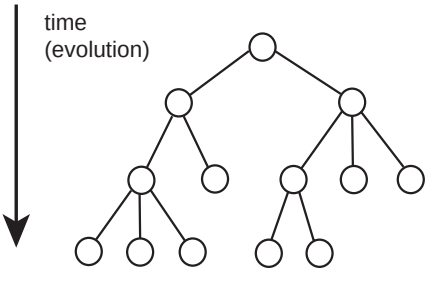

Figure 3. Phylogenetic tree.

The distance between two nodes can be measured by the length of the path that connects them. This means that even for discrete dimensions one can distinguish a rudimentary geometrical structure. For example, in the phylogenetic classification of animals that mirrors evolutionary branchings it is meaningful to say that rats and whales are more closely related than whales and fish.

3. Phenomenal and Scientific Interpretations of Dimensions

In order to separate different uses of quality dimensions it is important to introduce a distinction between a *phenomenal* (or *psychological*) and a *scientific* interpretation. The phenomenal interpretation concerns the cognitive structures (perceptions, memories, etc.) of humans or other organisms. The scientific interpretation, on the other hand, treats dimensions as a part of a scientific theory.

The distinction is relevant in relation to the two goals of cognitive science presented above. When the dimensions are seen as cognitive entities, that is, when the goal is to explain natural cognitive processes, their geometrical structure should not be determined by scientific theories which attempt at giving a "realistic" description of the world, but by *psychophysical* measurements which determine the structure of how our perceptions are represented. Furthermore, when it comes to providing a semantics for a natural language, the phenomenal interpretations of the quality dimensions are in focus.

On the other hand, when we are *constructing* an artificial system, the function of sensors, effectors and various control devices are in general described in terms of scientifically modeled dimensions. For example, the input variables of a robot may be a small number of physically measured magnitudes, like brightness, delay of a radar echo, or the pressure from a mechanical grip. With the aid of the programmed goals of the robot, these variables can then be transformed into a number of physical output magnitudes as, for example, the voltages of the motors controlling the left and the right wheels.

To give an example of the distinction, consider color. The distinction introduced here is supported by Gallistel (1990, p. 518–519) who writes:

> The facts about color vision suggest how deeply the nervous system may be committed to representing stimuli as points in descriptive spaces of modest dimensionality. It does this even for spectral compositions, which does not lend itself to such a representation. The resulting lack of correspondence between the psychological representation of spectral composition and spectral composition itself is a source of confusion and misunderstanding in scientific discussions of color. Scientists persist in referring to the physical characteristics of the stimulus and to the tuning characteristics of the transducers (the cones) as if psychological color terms like *red*, *green*, and *blue* had some straightforward translation into physical reality, when in fact they do not.

Gallistel's warning against confusion and misunderstanding of the two types of representation should be taken seriously. It is very easy to confound what science says about the characteristics of reality and our per-

ception of it. In this article, it is the phenomenal representation that will be in focus.

A *conceptual space* can now be defined as a collection of one or more quality dimensions. However, the dimensions of a conceptual space should not be seen as totally independent entities, but they are *correlated* in various ways since the properties of the objects modeled in the space covary. For example, the ripeness and the color dimensions covary in the space of fruits.

4. On the Origins of Quality Dimensions

In the previous sections, I have given some examples of quality dimensions from different kinds of domains. There seem to be different types of dimensions, so a warranted question is: Where do the dimensions come from? I do not believe there is a unique answer to this question. In this section, I will try to trace the origins of the different kinds of quality dimensions.

Firstly, some of the quality dimensions seem to be innate or developed very early in life. They are to some extent hardwired in our nervous system, as for example the sensory dimensions presented above. This probably also applies to our representations of ordinary space. Since domains of this kind are obviously extremely important for basic activities like finding food, avoiding danger, and getting around in the environment there is evolutionary justification for the innateness assumption. Humans and other animals who did not have a sufficiently adequate representation of the spatial structure of the external world were disadvantaged by natural selection.

The brain of humans and animals contains topographic areas mapping different kinds of sense modalities onto spatial areas. The structuring principles of these mappings are basically innate, even if the fine tuning is established during the development of the human or animal. The same principles seem to govern most of the animal kingdom. Gallistel (1990, p. 105) argues:

> [...] the intuitive belief that the cognitive maps of "lower" animals are weaker than our own is not well-founded. They may be impoverished relative to our own (have less on them) but they are not weaker in their formal characteristics. There is experimental evidence that even insect maps are metric maps.

Quine notes that something like innate quality dimensions is needed to make *learning* possible (Quine 1969, p. 123):

> Without some such prior spacing of qualities, we could never acquire a habit; all stimuli would be equally alike and equally differ-

> ent. These spacings of qualities, on the part of men and other animals, can be explored and mapped in the laboratory by experiments in conditioning and extinction. Needed as they are for all learning, these distinctive spacings cannot themselves all be learned; some must be innate.

However, once the process has started, new dimensions can be added by the learning process.[6] One kind of examples comes from studies of children's cognitive development. Two-year-olds can represent whole objects, but they cannot reason about the dimensions of the objects.

Learning new concepts is, consequently, often connected with *expanding* one's conceptual space with new quality dimensions. For example, consider the (phenomenal) dimension of *volume*. The experiments concerning "conservation" performed by Piaget and his followers indicate that small children have no separate mental dimension of volume; they confuse the volume of a liquid with the *height* of the liquid in its container. It is only at about an age of five years that they learn to represent the two dimensions separately. Similarly, three- and four-year-olds confuse *high* with *tall*, *big* with *bright*, etc (Carey 1978). Smith (1989, p. 146–147) argues that

> working out a system of perceptual dimension, a system of *kinds* of similarities, may be one of the major intellectual achievements of early childhood. [...] The basic developmental notion is one of differentiation, from global syncretic classes of perceptual resemblance and magnitude to dimensionally specific kinds of sameness and magnitude.

Still other dimensions may be *culturally* dependent.[7] Take "time" as an example: In some cultures time is conceived to be *circular* – the world keeps returning to the same point in time and the same events occur over and over again; and in other cultures it is hardly meaningful at all to speak of time as a dimension. A sophisticated time dimension, with the full metric structure, is needed for advanced forms of planning and coordination with other individuals, but is not necessary for the most basic activities of an organism. As a matter of fact, the standard Western conception of time is a comparatively recent phenomenon (see Toulmin and Goodfield 1965).

The examples given here indicate that many of the quality dimensions of human conceptual spaces are not directly generated from sensory inputs. This is even clearer when we use concepts based on the *functions*

[6] It must be noted that it is impossible to draw a sharp distinction between innate and learned quality dimensions, since many sensory dimensions are structurally prepared in the neural tissue at birth, but require exposure to sensory experiences to lay out the exact geometrical structure of the mapping.

[7] I do not claim that my typology of the origins of quality dimensions is exclusive, since, in a sense, all culturally dependent dimensions are also learned.

of artifacts or the *social roles* of people in a society. Even if we do not know much about the geometrical structures of these dimensions, it is quite obvious that there is some non-trivial such structure. This has been argued by Marr and Vaina (1982) and Vaina (1983), who give an analysis of functional representation where functions of an object are determined in terms of the actions it allows.

Culture, in the form of interaction between people, may in itself generate constraints on conceptual spaces. For example, Freyd (1983) puts forward the intriguing proposal that conceptual spaces may evolve as a representational form in a community just because people have to *share* knowledge (Freyd 1983, pp. 193–194):

> There have been a number of different approaches towards analyzing the structures in semantic domains, but what these approaches have in common is the goal of discovering constraints on knowledge representation. I argue that the structures the different semantic analyses uncover may stem from shareability constraints on knowledge representation. [...] So, if a set of terms can be shown to behave as if they are represented in a three-dimensional space, one inference that is often made is that there is both some psychological reality to the spatial reality (or some formally equivalent formulation) and some innate necessity to it. But it might be that the structural properties of the knowledge domain came about because such structural properties provide for the most efficient sharing of concepts. That is, we cannot be sure that the regularities tell us anything about how the brain can represent things, or even "prefer" to, if it didn't have to share concepts with other brains.

Here Freyd hints at an *economic* explanation of why we have conceptual spaces: they facilitate the sharing of knowledge.

Finally, some quality dimensions are introduced by *science*. Witness, for example, Newton's distinction between *weight* and *mass*, which is of crucial importance for the development of his mechanics, but which hardly has any correspondence in human perception. To the extent we have mental representations of the masses of objects in distinction to their weights, these are not given by the senses but have to be learned by adopting the conceptual space of Newtonian mechanics in our representations.

The most drastic changes in science occur when the underlying conceptual space is changed. I believe that most of the "paradigm shifts" discussed by Kuhn (1970) can be understood as *shifts of conceptual spaces*. I do not see any principal difference between this kind of change and the change involved in the development of a child's conceptual space. Introducing the distinction between "height" and "volume" is the same kind of phenomenon as when Newton introduced the distinction between "weight" and "mass". That distinction is nowadays ubiquitous in physics, even though there is only scant sensory support for it.

The conceptual space of Newtonian mechanics is, of course, based on scientific (theoretical) quality dimensions and not on phenomenal (psychological) dimensions. The quality dimensions of this theory are ordinary space (3-D Euclidean), time (isomorphic to the real numbers), mass (isomorphic to the non-negative real numbers), and force (3-D Euclidean space). Once a particle has been assigned a value for these eight dimensions, it is fully described as far as Newtonian mechanics is concerned. In this theory, an object is thus represented as a point in an 8-dimensional space.

5. Concept Formation Described with the Aid of Conceptual Spaces

In more abstract terms, a conceptual space S is established by a class of quality dimensions D_1, ..., D_n . A point in S is represented by a vector $v = < d_1, ..., d_n >$ with one index for each dimension. Each of the dimensions is endowed with a certain topological or metrical structure. The purpose of this section is to show how conceptual spaces can be used to model *concepts*.

A first rough idea is to describe a concept as a *region* of a conceptual space S, where "region" should be understood as a spatial notion determined by the topology and metric of S. For example, the point in the time dimension representing "now" divides this dimension, and thus the space of vectors, into two regions corresponding to the concepts "past" and "future". But the proposal suffers from a lack of precision as regards the notion of a "region". A more precise and powerful idea is the following criterion where the topological characteristics of the quality dimensions are utilized to introduce a spatial structure on concepts:

Criterion P:
A *natural concept* is a convex region of a conceptual space.

A *convex* region is characterized by the criterion that for every pair of points v_1 and v_2 in the region all points in between v_1 and v_2 are also in the region. The motivation for the criterion is that if some objects which are located at v_1 and v_2 in relation to some quality dimension (or several dimensions) both are examples of a concept C, then any object that is located between v_1 and v_2 on the quality dimension(s) will also be an example of C. I shall argue later that this criterion is psychologically realistic. It presumes that the notion of *betweenness* is meaningful for the relevant quality dimensions. This is a rather weak assumption which demands very little of the underlying topological structure.

Most concepts expressed by simple words in natural languages are natural concepts in the sense specified here. For instance, I conjecture that all *color terms* in natural languages express natural concepts with

respect to the psychological representation of the three color dimensions. In other words, the conjecture predicts that if some object o_1 is described by the color term C in a given language and another object o_2 is also said to have color C, then any object o_3 with a color that lies between the color of o_1 and that of o_2 will also be described by the color term C. It is well-known that different languages carve up the color circle in different ways, but all carvings seems to be done in terms of convex sets. Strong support for this conjecture can be found in Berlin and Kay (1969), although they do not treat color terms in general but concentrate on basic color terms. On the other hand, the reference of an artificial color term like "grue" (Goodman 1955) will not be a convex region in the ordinary conceptual space and thus it is not a natural concept according to Criterion P.[8]

Another illustration of how the convexity of regions determines concepts and categorizations is the phonetic identification of *vowels* in various languages. According to phonetic theory, what determines a vowel are the relations between the basic frequency of the sound and its formants (higher frequencies that are present at the same time). In general, the first two formants F_1 and F_2 are sufficient to identify a vowel. This means that the coordinates of two-dimensional space spanned by F_1 and F_2 (in relation to a fixed basic pitch F_0) can be used as a fairly accurate description of a vowel. Fairbanks and Grubb (1961) investigated how people produce and recognize vowels in "General American" speech.

Figure 4 summarizes some of their findings. The scale of the abscissa and ordinate are the logarithm of the frequencies of F_1 and F_2 (the basic frequency of the vowels was 130 Hz). As can be seen from the diagram, the preferred, identified and self-approved examples of different vowels form convex subregions of the space determined by F_1 and F_2 with the given scales.[9] As in the case of color terms, different languages carve up the phonetic space in different ways (the number of vowels identified in different languages varies considerably), but I conjecture again that each vowel in a language will correspond to a convex region of the formant space.

An important thing to note in this example is that identifying F_1 and F_2 as the relevant dimensions for vowel formation is a phonetic *discovery.* We had the concepts of vowels already before this discovery, but the spatial analysis makes it possible for us to understand several features of the classifications of vowels in different languages.

Criterion P provides an account of concepts that is independent of

[8] For an extended analysis of this example see Gärdenfors (1989).

[9] A *self-approved* vowel is one that was produced by the speaker and later approved of as an example of the intended kind. An *identified* sample of a vowel is one that was correctly identified by 75% of the observers. The *preferred* samples of a vowel are those which are "the most representative samples from among the most readily identified samples" (Fairbanks and Grubb 1961, p. 210).

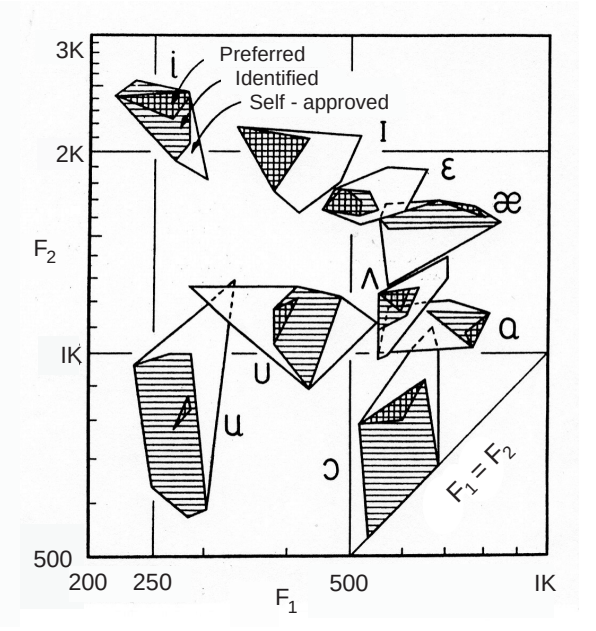

Figure 4: Frequency ranges of different vowels in the two-dimensional space generated by the first two formants (from Fairbanks and Grubb 1961).

both possible worlds and individuals and it satisfies Stalnaker's *desideratum* that a concept "... must be not just a rule for grouping individuals, but a feature of individuals in virtue of which they may be grouped" (Stalnaker 1981, p. 347). However it should be emphasized that I only view the criterion as a *necessary* but perhaps not sufficient condition on a natural concept. The criterion delimits the class of concepts that are useful for cognitive purposes, but it may not be sufficiently restrictive.

6. Relations to Prototype Theory

Describing concepts as convex regions of conceptual spaces fits very well with the so called *prototype theory* of categorization developed by Rosch and her collaborators (Rosch 1975, 1978, Mervis and Rosch 1981, Lakoff 1987). The main idea of prototype theory is that within a category of objects, like those instantiating a concept, certain members are judged to be more representative of the category than others. For example, robins are judged to be more representative of the category "bird" than are ravens, penguins and emus; and desk chairs are more typical instances of the category "chair" than rocking chairs, deck-chairs, and beanbag chairs. The most representative members of a category are called *prototypical*

members. It is well-known that some concepts, like "red" and "bald" have no sharp boundaries and for these it is perhaps not surprising that one finds prototypical effects. However, these effects have been found for most concepts including those with comparatively clear boundaries like "bird" and "chair".

In traditional philosophical analyses of concepts, based on truth-functions or possible worlds it is very difficult to explain such prototype effects (see Gärdenfors 1991). Either an object is a member of the class assigned to a concept (relative to a given possible world) or it is not and all members of the class have equal status as category members. Rosch's research has been aimed at showing asymmetries among category members and asymmetric structures within categories. Since the traditional definition of a concept neither predicts nor explains such asymmetries, something else must be going on.

In contrast, if concepts are described as convex regions of a conceptual space, prototype effects are indeed to be expected. In a convex region one can describe positions as being more or less *central*. For example, if color concepts are identified with convex subsets of the color space, the central points of these regions would be the most prototypical examples of the color. In a series of experiments, Rosch has been able to demonstrate the psychological reality of such "focal" colors. For another illustration we can return to the categorization of vowels presented in the previous section. Here the structure of the subjects' different kinds of responses shows clear prototype effects.

For more complex categories like "bird" it is perhaps more difficult to describe the underlying conceptual space. However, if something like the analysis of shapes by Marr and Nishihara (1978) is adopted, we can begin to see how such a space would appear.[10] Their scheme for describing biological forms uses hierarchies of cylinder-like modeling primitives. Each cylinder is described by two coordinates (length and width). Cylinders are combined by determining the angle between the dominating cylinder and the added one (two polar coordinates) and the position of the added cylinder in relation to the dominating one (two coordinates). The details of the representation are not important in the present context, but it is worth noting that on each level of the hierarchy an object is described by a comparatively small number of coordinates based on lengths and angles. Thus the object can be identified as a hierarchically structured vector in a (higher order) conceptual space. Figure 5 provides an illustration of this hierarchical structure.

It should be noted that even if different members of a category are judged to be more or less prototypical, it does not follow that some of the

[10]This analysis is expanded in Marr (1982), Chap. 5. A related model, together with some psychological grounding, is presented by Biederman (1987).

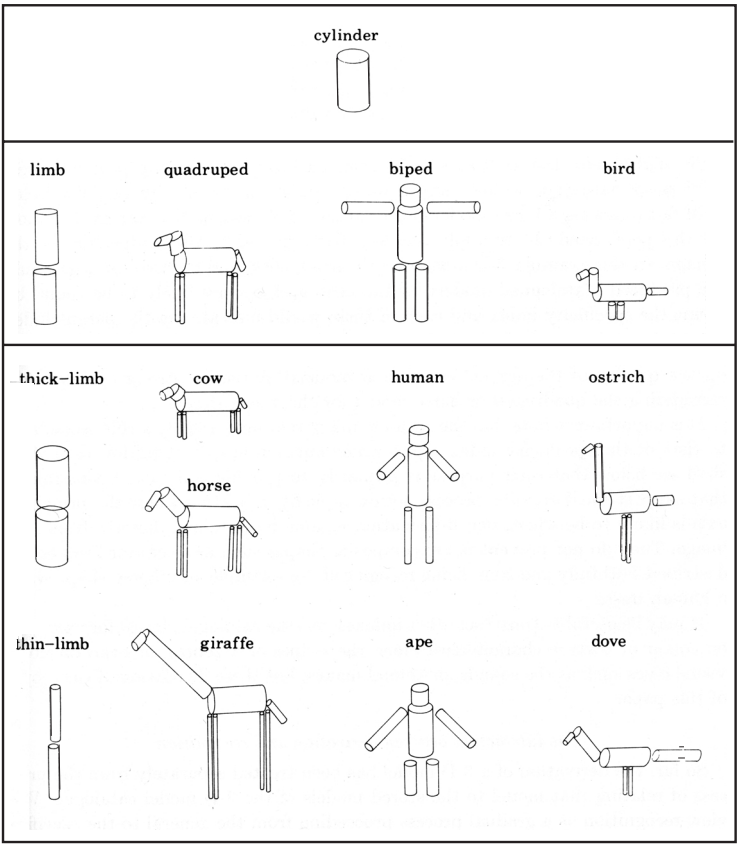

Figure 5: Representing shapes by cylinders (from Marr and Nishihara 1978).

existing objects must represent "the prototype". If a concept is viewed as a convex region of a conceptual space this is easily explained, since the central member of the region (if unique) is a possible individual in the sense discussed above (if all its dimensions are specified) but need not be among the existing members of the category. Such a prototype point in the region need not be completely described as an individual, but is normally represented as a partial vector, where only the values of the dimensions that are relevant to the concept have been determined. For example, the general shape of the prototypical bird would be included in the vector, but its color or age presumably would not.

It is possible to argue in the converse direction, too, and show that, if prototype theory is adopted, then the representation of concepts as convex regions is to be expected. Assume that some quality dimensions of a conceptual space are given, for example the dimensions of color space,

and that we want to partition it into a number of categories, for example color categories. If we start from a set of prototypes p_1, ..., p_n of the categories, for example the focal colors, then these should be the central points in the categories they represent. One way of using this information is to assume that for every point p in the space one can measure the *distance* from p to each of the p_i's. If we now stipulate that p belongs to the same category as the *closest* prototype p_i, it can be shown that this rule will generate a partitioning of the space that *consists of convex areas* (convexity is here defined in terms of an assumed distance measure). This is the so-called *Voronoi tessellation*, a two-dimensional example of which is illustrated in Figure 6.

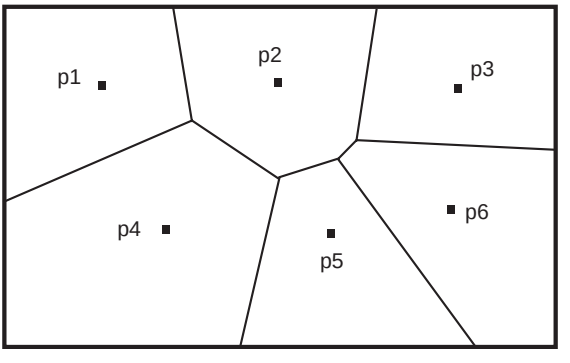

Figure 6: Voronoi tessellation of the plane into convex sets.

Thus, assuming that a metric is defined on the subspace that is subject to categorization, a set of prototypes will by this method generate a unique partitioning of the subspace into convex regions. Hence there is an intimate link between prototype theory and the description of concepts as convex regions in a conceptual space.

As a concrete instance of this technique, Petitot (1989) applied Voronoi categorization to explain some aspects related to the categorical perception of phonemes. In particular, he analyzed the relations between the so called stop consonants /b/, /d/, /g/, /p/, /t/, /k/. The relations between these consonants are expressed with the aid of two dimensions: one is the voiced-unvoiced dimension, the other is the labial-dental-velar dimension which relates to the place of articulation of the consonant. Both these dimension can be treated as continuous. Figure 7 shows how he represents the boundaries between the six consonants.

As an example of the information contained in this model, Petitot points out (Petitot 1989, p. 68):

> The *geometry* of the system of boundaries can provide precious information about the hierarchical relations that stop consonants

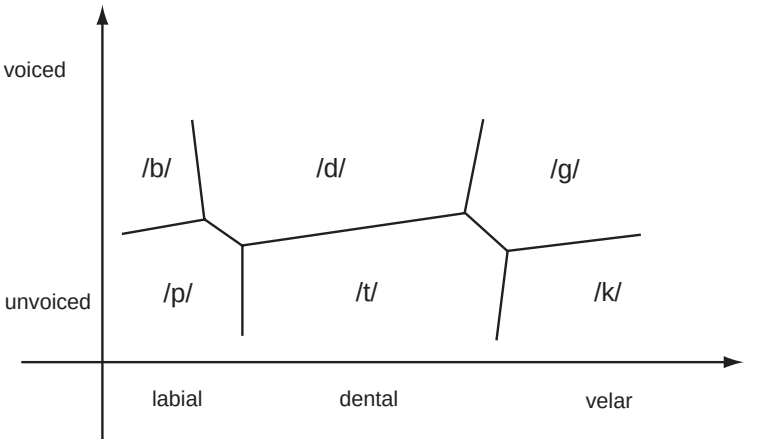

Figure 7: A Voronoi model of the boundaries of stop consonants (from Petitot (1989), p. 69).

maintain with each other. The fact that in the model of Massaro and Oden, the domains of /p/ and /d/ are *adjacent*, whereas those of /b/ and /t/ are *separated*, indicates that the contrast between /b/ and /t/ is much greater than that between /p/ and /d/.

7. Conclusions

The main purpose of this article is to present the core of the theory of conceptual spaces. In this connection an important question is: what *kind* of theory is the theory of conceptual spaces? Is it an empirical, normative, computational, psychological, neuroscientific, or linguistic theory?

As was stated in Sect. 1, cognitive science has two predominant goals: to *explain* cognitive phenomena and to *construct* artificial systems that can solve various cognitive tasks. The theory of conceptual spaces is presented as a *framework for representing information*. It should be seen as a theory that complements the symbolic and the connectionist models and forms a bridge between these forms of representation.

The primary aim is to use the theory of conceptual spaces in constructive tasks. In previous work, I have shown how it can be used in computational models of *concept formation* (Gärdenfors 1992) and *induction* (Gärdenfors 1990, 1993) and that it is useful for representing the *meanings* of different kinds of linguistic expressions in a computational approach to semantics.

The borderline between constructive and explanatory uses of conceptual spaces is not sharp. When, for example, constructing the representational world of a robot, it is often worthwhile to take lessons from how

biology has solved the problems in the brains of humans and other animals. Conversely, the construction of an artificial system that can successfully solve a particular cognitive problem may provide clues to how an empirical investigation of biological systems should proceed. Consequently, there is a spiraling interaction between constructive and explanatory uses of conceptual spaces.

This article has been asking questions about the geometry of thought. With the aid of the notion of conceptual spaces I have provided an analysis of concepts. A key notion is that of a *natural concept* which is defined in terms of well-behaved regions of conceptual spaces – a definition that crucially involves the geometrical structure of the various domains.

In my opinion, a conceptual level of representation should play a central role in the cognitives sciences. After having been dominant for many years, the symbolic approach was challenged by connectionism (which is nowadays broadened to a wider study of dynamical systems). However, for many purposes the symbolic level of representation is too coarse, and the connectionist too fine-grained. In relation to the two goals of cognitive science, I submit that the conceptual level will add significantly to our explanatory capacities when it comes to understanding cognitive processes, in particular those connected with concept formation and language understanding.

Where do we go from here? The main factor preventing a rapid development of different applications of conceptual spaces is the lack of knowledge about the relevant quality dimensions. It is almost only for perceptual dimensions that psychophysical research has succeeded in identifying the underlying geometrical and topological structures (and, in rare cases, the psychological metric). For example, we only have a very sketchy understanding of how we perceive and conceptualize things according to their shapes.

When the structure of the dimensions of a particular domain is discovered, this often leads to fruitful research. For example, the development of the vowel space that was presented in section 5 led to a wealth of new results in phonetics and a deeper understanding of the speech process.

Thus, those who want to contribute to the research program should start hunting for the hidden conceptual spaces. Even if results may not be easily forthcoming, they are sure to have repercussions in other areas of cognitive science as well.

References

Berlin B. and Kay P. (1969): *Basic Color Terms: Their Universality and Evolution*, University of California Press, Berkeley.

Biederman I. (1987): Recognition-by-components: a theory of human image understanding. *Psychological Review* **94**, 115–147.

Clark A. (1993): *Sensory Qualities*, Clarendon Press, Oxford.

Fairbanks G. and Grubb P. (1961): A psychophysical investigation of vowel formants. *Journal of Speech and Hearing Research* **4**, 203–219.

Freyd J. (1983): Shareability: the social psychology of epistemology. *Cognitive Science* **7**, 191–210.

Gabrielsson A. (1981): Music psychology – a survey of problems and current research activities. In *Basic Musical Functions and Musical Ability*, Publications of the Royal Swedish Academy of Music, No. 32, pp. 7–80.

Gallistel C.R. (1990): *The Organization of Learning*, MIT Press, Cambridge, MA.

Gärdenfors P. (1988): Semantics, conceptual spaces and music. In *Essays on the Philosophy of Music* (*Acta Philosophica Fennica, vol. 43*), ed. by V. Rantala, L. Rowell and E. Tarasti, The Philosophical Society of Finland, Helsinki, pp. 9–27.

Gärdenfors P. (1990): Induction, conceptual spaces and AI. *Philosophy of Science* **57**, 78–95.

Gärdenfors P. (1991): Frameworks for properties: possible worlds vs. conceptual spaces. In *Language, Knowledge and Intentionality (Acta Philosophica Fennica, vol. 49)*, ed. by L. Haaparanta, M. Kusch, and I. Niiniluoto, The Philosophical Society of Finland, Helsinki, pp. 383–407.

Gärdenfors P. (1993): Induction and the evolution of conceptual spaces. In *Charles S. Peirce and the Philosophy of Science*, ed. by E.C. Moore, University of Alabama Press, Tuscaloosa, pp. 72–88.

Gärdenfors P. (2000): *Conceptual Spaces: On the Geometry of Thought*, MIT Press, Cambridge, MA.

Goodman N. (1955): *Fact, Fiction and Forecast*, Harvard University Press, Cambridge, MA.

Henning H. (1916): Die Qualitätenreihe des Geschmacks. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* **74**, 203–219.

Kuhn T. (1970): *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

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

Marr D. (1982): *Vision*, Freeman, San Francisco.

Marr D. and Nishihara H.K. (1978): Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society London B* **200**, 269–294.

Marr D. and Vaina L. (1982): Representation and recognition of the movements of shapes. *Proceedings of the Royal Society London B* **214**, 501–524.

Mervis C. and Rosch E. (1981): Categorization of natural objects. *Annual Review of Psychology* **32**, 89–115.

Petitot J. (1989): Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics* **15**, 25–71.

Quine W.V.O. (1969): Natural kinds. In *Ontological Relativity and Other Essays*, Columbia University Press, New York, pp. 114–138.

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

Rosch E. (1978): Prototype classification and logical classification: the two systems. In *New Trends in Cognitive Representation: Challenges to Piaget's Theory*, ed. by E. Scholnik, Lawrence Erlbaum, Hillsdale, pp. 73–86.

Schiffman H.R. (1982): *Sensation and Perception*, Wiley, New York.

Smith L.B. (1989): From global similarities to kinds of similarities – the construction of dimensions in development. In *Similarity and Analogical Reasoning*, ed. by S. Vosniadou and A. Ortony, Cambridge University Press, Cambridge, pp. 146–178.

Stalnaker R. (1981): Antiessentialism. *Midwest Studies of Philosophy* **4**, 343–355.

Toulmin S. and Goodfield J. (1965): *The Discovery of Time*, Penguin, Harmondsworth.

Vaina L. (1983): From shapes and movements to objects and actions. *Synthese* **54**, 3–36.