Chapter 2
Conceptual Spaces, Generalisation Probabilities and Perceptual Categorisation

Nina L. Poth

Abstract Shepard's (Science 237(4820):1317–1323, 1987) *universal law of generalisation* (ULG) illustrates that an invariant gradient of generalisation across species and across stimuli conditions can be obtained by mapping the probability of a generalisation response onto the representations of similarity between individual stimuli. Tenenbaum and Griffiths (Behav Brain Sci 24:629–640, 2001) Bayesian account of generalisation expands ULG towards generalisation from multiple examples. Though the Bayesian model starts from Shepard's account it refrains from any commitment to the notion of psychological similarity to explain categorisation. This chapter presents the conceptual spaces theory as a mediator between Shepard's and Tenenbaum & Griffiths' conflicting views on the role of psychological similarity for a successful model of categorisation. It suggests that the conceptual spaces theory can help to improve the Bayesian model while finding an explanatory role for psychological similarity.

2.1 Introduction

As a counter to the behaviouristically inspired idea that generalisation of a particular kind of behaviour from one single stimulus to another single stimulus is a mere failure of discrimination, Shepard (1987) formulated a law that he empirically demonstrated to obtain across stimuli and species. His argument was that the law models categorisation as a cognitive function of perceived similarities.

The ULG has contributed to many models in categorisation research. One such a model that evolved on the basis of his work is Tenenbaum and Griffiths'(from herein T&G, 2001) Bayesian inference model of categorisation. T&G argue that

N. L. Poth (✉)
School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK
e-mail: nina.poth@ed.ac.uk

M. Kaipainen et al. (eds.), *Conceptual Spaces: Elaborations and Applications*, Synthese Library 405, https://doi.org/10.1007/978-3-030-12800-5_2

their model is advantageous to Shepard's model in two respects. On the one hand, it can capture the influence of multiple examples on categorisation behaviour. On the other hand, T&G argue that it can unify two previously incompatible approaches to similarity. One is Shepard's approach to similarity as a function of continuous distance in multi-dimensional psychological space. The other is Tversky's (1977) set-theoretic model of similarity, which considers the similarity of two items to be a function of the number of their shared or distinct features. T&G argue that their model is advantageous to Shepard's original proposal because it is formally compatible with both conceptions of similarity and thus scores high in terms of its unificatory power. However, T&G take as an implication of the fact that their model is not strictly committed to any particular conception of similarity (i.e. Shepard's or Tversky's), that the (scientific) concept of similarity can be generally dismissed from explanations of the universal gradient of generalisation that Shepard had observed; probabilities alone are sufficient. Contra Shepard, T&G thus suggest considering generalisation probabilities as primary to similarity.

In this chapter, I suggest that the theory of conceptual spaces offers a perfect tool for resolving this debate. In particular, I argue that the theory of conceptual spaces can make T&G's Bayesian model more conceptually transparent and psychologically plausible by offering a tool to supplement it with a psychological similarity space, while capturing its advantage of showing that the multiplicity of examples in a learner's history matters for changes in categorisation behaviour. The conceptual spaces theory then helps to explicate that some notion of similarity is indeed needed for probabilistic models of categorisation more generally, and hence keeps in with Shepard's original motivation to explain categorisation as a function of perceptual similarity.

In Sect. 2.2, I outline Shepard's (1987) model of categorisation, with an emphasis on the role he attributes to perceived similarities in categorisation. In Sect. 2.3, I present T&G's (2001) expansion of Shepard's model, with an emphasis on the size principle – a principle which formally expresses the added value from considering multiple examples for categorisation. In Sect. 2.4, I present some problems with T&G's model. I argue that T&G's conclusion that probabilities should be considered primary to similarities is not warranted and that this perspective undermines their model's semantic interpretability. In Sect. 2.5 I suggest, more positively and on the basis of Decock and colleagues' (2016) Geometric Principle of Indifference, to consider a conceptual space as a semantic basis for Bayesian hypotheses spaces. I argue that in providing such a space, the conceptual spaces theory can help to avoid the issues with T&G's model and bring it in line with Shepard's original motivation to explain the generalisation gradient as a psychological function of similarity. I conclude with a suggestion, that this combination of a conceptual space and Bayesian inference could be considered as a more fruitful approach to modelling generalisation probabilities in perceptual categorisation than a probabilistic model on its own.

2.2 Shepard's Universal Law of Generalisation

This section briefly reviews Shepard's mathematical model of generalisation, and points out its historical relevancy for cognitive-representational models of categorisation.

Shepard (1987)'s universal law of generalisation is a pioneering concept in the psychology of perceptual categorisation. When Shepard published the results of his work, it was widely held that categorisation does not follow a universal law or pattern that might reflect a natural kind structure. Shepard contrasts the case of categorisation to Newton's (1687/1934) universal law of gravitation. Newton's law was very influential and helped to discover invariances in the physical structure of the universe. Inspired by Newton's achievements, Shepard's aim was to find a mathematical generalisation function that accurately models the psychological representation of cognitive categories by extracting the invariances in the perceived members of a category. Shepard took this to be a vital move against behaviourism and for the idea that generalisation is a cognitive decision, not merely a failure of sensory discrimination.

Shepard's law can be expressed by the following proposition.

(ULG) The universal law of generalisation:
For a set of stimuli, i and j, the empirical probability of an organism to generalise a type of behaviour towards j upon having observed i is a monotone and exponentially decreasing function of the distance between i and j in a continuous psychological similarity space.

ULG states that with a continuous increase in the distance between stimuli i and j in psychological space (that is, with an increase in their perceived dissimilarity), subjects are decreasingly likely to give these stimuli the same behavioural response. On this basis, ULG predicts that subjects should be less likely to generalise a behaviour associated with a given physical stimulus towards a relatively dissimilar stimulus and more likely to generalise the behaviour towards a relatively similar stimulus.

Shepard captured this tendency formally in terms of an exponential decay function. He obtained this function by plotting the probability of generalisation (the observed relative frequency at which a subject generalises behaviour to stimulus i towards stimulus j), g_{ij}, against a measure of psychological stimulus distance, d_{ij}, where psychological distance was obtained by means of the multi-dimensional scaling method (Carroll and Arabie 1980; Kruskal 1964; Shepard 1962). Shepard showed that the generalisation gradient is invariant across various stimuli (e.g. size, lightness & saturation, spectral hues, vowel and consonant phonemes and Morse code signals) and across species (e.g. pigeons and humans), thus the name 'universal law'. He obtained two insights from the mathematical modelling of this law.

1. If measured based on psychological (instead of physical) distance, the shape of the generalisation gradient is uniform across stimuli and species. (Uniformity)
2. The metric of the psychological similarity space is either the City-Block distance or the Euclidean metric. (L_1-/L_2-measurability)

The first point expresses the idea that differences in stimulus strength and corresponding generalisation might depend on differences in the psychophysical function that transforms physical measurements into psychological ones. For example, subjects might generalise the same label to two colour shades that a physicist would classify as 'green' and 'yellow' along the physical wavelength spectrum. However, if the two colour shades are positioned in a model of psychological colour space instead[1] then subjects' generalisations might be expected. This is because in psychological similarity space, the colour shades may be judged to be more similar than their measure along the physical wavelength spectrum actually indicates. The physical distance between stimuli along the one-dimensional wavelength spectrum might thus differ from their perceived distance in multi-dimensional psychological similarity space. Shepard took this discrepancy as a possible explanation of the previous difficulty to establish an empirically adequate model of generalisation by measuring physical stimulus space. Thus, a transformation function would be needed to recover a psychological distance measure from the physical distance measure. Shepard's second insight was that this function is a member of the family of Minkowski metric.

Categories in this psychological framework are modelled as consequential regions in multidimensional similarity space. Shepard assumes three constraints on the categoriser's background beliefs about consequential regions prior to any observation.

> ... (i) all locations are equally probable; (ii) the probability that the region has size s is given by a density function $p(s)$ with finite expectation μ; and (iii) the region is convex, of finite extension in all directions, and centrally symmetric. (Shepard 1987, 1320)

(i) is important because it yet does not assume that there are differences in the internal structure of categories. This is relevant because if any possible item in a category has the same chance of occurring, then this model cannot account for prototype effects in categorisation (cf. Rosch and Mervis 1975). (ii) is advantageous with respect to the formal precision and flexibility of the model. Since the magnitudes of the measured stimuli (e.g. brightness and sound) are measured in continuous space, probability densities are a suitable tool to use for a decision strategy when evaluating candidate categories on the basis of the training stimulus. (iii) is an assumption that makes the model mathematically more elegant, but Shepard has given additional arguments for assuming that categorisers indeed categorise in ways that satisfy convexity (Shepard 1981).

[1] A common example is the 3-dimensional CIELAB colour space which models colour representations along three axes, the hue, saturation and brightness dimensions (cf. Fairchild 2013).

As a psychophysical law, ULG explains the probability of generalisation by heavily relying on the notion of internally represented similarities. What makes the generalisation function psychological as opposed to physical is that "it can be determined in the absence of any physical measurements on the stimuli." (Shepard 1987, 1318). For example, even if the colour of an item on the physical wavelength spectrum might change so as to become more different under differing lighting conditions, this change might not actually be represented as a change in the vector coordinates that would be assigned to the perceived colour of the item in psychological similarity space. Thus, the invariance of the gradient could not be explained without the more subjective notion of representational perceptual similarity.

If this is correct, and the generalisation gradient arises from psychological instead of physical measurements, then what is needed for a model of similarity-based categorisation is a conceptual distinction between the psychological and physical magnitude of the difference between (training- and test-) stimuli, respectively, how they are related. In his work on psychophysical complementarity, Shepard (1981) argues that psychological similarity offers such a distinction. In particular, Shepard distinguishes between two kinds of similarities; first- and second-order similarities. Accordingly, first-order similarities are similarities between physically measurable properties in the world on the one hand and representations thereof on the other hand. For example, consider the similarity between the redness of a dress as measured on the physical wavelength spectrum, and the redness of the dress as I perceive it. Second-order similarities, in contrast, are similarities between mental representations themselves. For instance, consider the similarity between my representation of the dress' redness at one point in time, t_0, as compared to my perceptual experience of the dress' redness at another point in time, t_1.

Why should it be important to distinguish between first- and second-order similarities for categorisation? Because they impose different kinds of accuracy conditions on the representations of similarities, which in turn constitute the generalisation gradient. Edelman (1998) motivates this point by alluding to Shepard & Chipman's (1970) distinction between first- and second-order isomorphisms.[2] They suggest that veridicality in perception is instantiated through the perception of similarities amongst the structure of shapes. The task of perceptual categorisation is not to build representations that resemble objects in the world. Instead, the task of the visual system is to build representations that stand in some orderly relationship to similarity relations between *perceived* objects. This supports Shepard's idea that the criteria for whether generalisation is accurate or not are not determined by physical measurements but by some psychological standard.

[2]For instance, second-order isomorphisms of shapes are measurements of similarities between representations of the similarities of shapes (Edelman refers to this as a 'representation of similarity'), as opposed to similarities between distal shapes and their proximal representations (a 'representation by similarity'). Where categories are seen here as reference shapes, they are to some extent cognitive constructions.

Consider second-order isomorphisms between shapes. What is represented are principled quantifications of changes of shape, not shapes themselves. The idea is that information about distal (represented) similarity relationships is picked up by the representing system through a translation process. In particular, the information about similarity relationships is reduced to invariances between movements in distal parameter space. This allows for dimensionality reduction to constitute a proximal (representing) shape space (Edelman 1998, 452). This process of translation (as opposed to a process of reconstruction), allows for a reverse inference from subjects' similarity judgements to a common metric. Veridicality, in that sense, means consistency amongst subjects when judging the similarities between considered object shapes, as opposed to consistency across stimuli conditions (for individual shapes).

The example of first- and second-order similarities illustrates that Shepard considers psychological similarity as explanatorily central to the relationship between assignments of category membership and perceived similarity. This goes against behaviourist analyses because similarity is seen as a cognitive function of decreasing distance. But Shepard's model is restricted to a comparison between representations of single members of a category. An alternative view on generalisation probabilities is offered by a Bayesian model of categorisation that considers generalisation from multiple examples but eventually suggests to explain categorisation regardless of the notion of psychological similarity. The model is presented in the next section.

2.3 Tenenbaum and Griffiths' Size Principle

This section outlines a Bayesian model of categorisation by Tenenbaum and Griffiths (2001) that attempts to expand Shepard's approach to generalisation in two ways.

1. They show that the number and magnitude of examples observed shapes the generalisation gradient.
2. They show that the probability of generalisation is (formally) independent of any particular model of similarity.

I elaborate shortly on both points to illustrate the differences between T&G's and Shepard's views on the relation between generalisation probabilities and psychological similarities.

The first point of expansion considers the generalisation function that learners are supposed to follow when learning categories. For this, T&G suggest a Bayesian inference algorithm, which they call the *size principle*. It helps to consider the size principle in light of the general Bayesian learning theory that T&G suggest.

The idea is that learners follow Bayes' theorem in computing the posterior probability, $Pr(H|E)$, of a hypothesis, H, about which consequential region is shared for stimuli of a common class, in light of the available evidence, E.

Bayes' Theorem

$$Pr(H|E) = \frac{Pr(H)Pr(E|H)}{Pr(E)}$$

Bayes' Theorem makes explicit how the posterior probability of a hypothesis given some piece of evidence can be obtained; by taking the prior probability, $Pr(H)$, together with the likelihood, $Pr(E|H)$, relative to the probability of the evidence, $Pr(E)$. For the current purpose, only the prior and likelihood are of interest. This is because dividing by $Pr(E)$ only serves normalisation purposes.

T&G argue that the likelihood term can be replaced by the size principle. The size principle states that if the available evidence is held constant, hypotheses that point towards smaller consequential regions should be preferred over those hypotheses that suggest larger consequential regions when making a generalisation decision. Moreover, if the information about perceived similarities is held constant, the tendency to prefer smaller categories for generalisation should become stronger with an increasing number of examples observed for that category. Formally, the size principle can be expressed as follows.

The size principle

$$Pr(E|H) \propto \left(\frac{1}{size(H_C)}\right)^{|n|} \tag{2.1}$$

Consider an example for the size principle that comes from Xu and Tenenbaum (2007). Three Dalmatians are given as examples for the word 'fep', together with the following hypotheses space.

H_1: $\langle$'fep', DALMATIAN$\rangle$
H_2: $\langle$'fep', DOG$\rangle$
H_3: $\langle$'fep', WHITE WITH BLACK DOTS$\rangle$

If the three Dalmatians are a random sample of the true category that the word 'fep' refers to, the size principle says that learners should have a higher degree of belief in H_1 than in H_2 and H_3. Following this size principle, this is a rational choice because it is more likely to observe 3 Dalmatians as examples of what 'fep' means if in fact it referred to the category of Dalmatians as compared to the class of dogs or things that are white and have black dots. The size principle thereby expresses what Xu and Tenenbaum (2007) call a *suspicious coincidence* mechanism: it would be very unlikely to observe 3 Dalmatians if 'fep' meant 'dog'. More formally, based on the size principle: since $size(Dalmatian) < size(dog) < size(white\ with\ black\ dots)$, $Pr(E|H_1) > Pr(E|H_2) > Pr(E|H_3)$.

The second point of T&G's expansion takes the size principle for granted and argues that the probability of generalisation – expressed in the likelihood term – is itself primary to perceived similarities in the analysis of categorisation behaviour.

This claim is motivated by the idea that formally, the probabilistic model can be translated into either a set-theoretic approach to similarity (such as the one put forward by Tversky 1977) or a similarity-as-psychological-distance approach (such as the one suggested by Shepard 1981, 1987). The probability space that is considered in T&G's model can be measured in terms of both, the number of shared or distinct features or the relative distances between items in a continuous space. This is because relative frequencies can be expressed in terms of either spatial or numeral proportions. Thus, based on which quantity the size of a consequential region is actually measured is irrelevant to Eq. 2.1.

T&G take the formal independence of probabilities from the particular type of similarity model to indicate that perceptual similarities might be derived from generalisation probabilities. T&G conclude that generalisation probabilities are enough to explain categorisation behaviour and that such explanations can therefore do without the concept of psychological similarity. They write:

> We expect that, depending on the context of judgement, the similarity of y to x may involve the probability of generalizing from x to y, or from y to x or some combination of those two. It may also depend on other factors altogether. Qualifications aside, interesting consequences nonetheless follow just from the hypothesis that similarity somehow depends on generalization, without specifying the exact nature of the dependence. (Tenenbaum and Griffiths 2001, 637)

However, it is yet an open question how probability assignments can determine perceived similarities between category members. More on this problem in Sect. 2.4.

2.3.1 Advantages Over Shepard's Model

The size principle adds to Shepard's model in two ways. One advantage is that T&G's model makes the role of exemplar variability for generalisation more transparent. This helps understanding why multiple examples, expressed by the exponent $|n|$, can help to make categorisation more precise. Accordingly, a higher number of examples helps to gain more information about where to set the boundaries of a category: "All other things being equal, the more examples observed within a given range, the lower the probability of generalization outside that range." (Tenenbaum and Griffiths 2001, 633). Categorisation is then a form of prediction. A category plays the role of a random variable, $X = \{x_1, \ldots, x_n\}$ whose probability to take on particular values in similarity space must be calculated. A more informative distribution of the estimated random variable allows for more precise predictions which values in similarity space are most probably to be observed as next examples of the to-be-inferred category.

Another advantage of T&G's model over Shepard's model is that it is richer in the predictions that it makes about generalisation performance. This also follows from the additional consideration of exemplar variability in the size principle. If less data is available under equal exemplar variability, the uncertainty in generalisation is greater and the overall distribution of probability assignments less distinct. Holding the exponent stable, the size principle prescribes a preference for categories with a smaller magnitude. Exemplar-variability, accordingly, determines generalisation probability. To that end, T&G can explain changes in the generalisation gradient that occur despite constancies in similarity comparisons. For instance, the model predicts that even under circumstances in which a novel item is actually very similar to the average of previously observed exemplars, it might not be generalised upon. The rationale is given by the size principle. With an increasing amount of examples, rational learners should become more restrictive in their willingness to expand generalisation. In other words, boundaries around regions in similarity space become sharper and generalisation patterns more distinct throughout a subject's learning history. Thereby, the Bayesian model can account for undergeneralisation – some of the presented examples are not considered particularly relevant for generalisation, even though they would fall within the hypothesised consequential region. Since this result is difficult to obtain in Shepard's model, the size principle offers a valuable expansion of ULG.

2.4 Problems for Tenenbaum and Griffiths' Approach

This section outlines three problems for T&G's approach. The first problem is that T&G are too hasty in dismissing similarity from their programme of explaining generalisation. T&G illustrate that their Bayesian model is independent of any particular account of similarity. However, T&G conclude that just because the Bayesian model is formally independent of any particular notion of similarity, similarity can be dismissed from an explanation of the generalisation gradient. They even make the stronger claim that the represented generalisation probabilities should be seen as primary to the perceived similarities of category members. But this conclusion seems too hasty. Just because their model is independent of any particular view on similarity (i.e. it is formally compatible with both Tversky's and Shepard's definition of similarity), this does not mean that similarity more generally cannot serve as a conceptually useful notion for the probabilistic model. Indeed, it could be plausible to think that considerations of similarity would, in fact, make T&G's Bayesian inference approach more precise and explanatorily useful. The first problem has two facets in effect.

The first facet of the problem of disregarding similarity is that similarity is conducive to an explication of the relationship between exemplar variability and how this changes the generalisation gradient. The reason for this is that it is hard to see how the concept of exemplar variability itself can be defined without alluding to similarity in the input data; in fact, it is easier to see how generalisation probabilities can be derived from changes in the average similarity between examples. This also corresponds to the common conception of exemplar variability in the literature on clustering algorithms in machine learning, particularly on nearest-neighbour models (Russell and Norvig 2002, ch. 18). The remaining question for T&G's model is thus which of its features can establish the relationship between the observed generalisation gradient and the agent's corresponding subjective degrees of belief about category membership, if not similarity.

The second facet is that T&G underestimate the semantic value of similarity for interpreting the size principle. Yet Eq. 2.1 leaves open how the size of a candidate category should be measured, and its semantic content interpreted. One reason for its semantic opaqueness is that Eq. 2.1 loses the explicit mentioning of the evidence on the right-hand side of the equation.[3] This also makes it difficult to identify the degree of confirmation of a belief by an example because it cannot be compared how far and by which measure the contents of terms E and H align. More conceptually, it is not clear what the belief in the Bayesian model that is assigned a probability value is actually about. Similarity could provide an answer to this question. For instance, on Shepard's account, consequential regions are individuated through a measure of distance. This measure is an internal representation of the agent, and thus provides content for a belief about category membership. Similarity can help understanding of the relationship between the perceptual features of observations as represented by the agent and the probabilistic inference of concepts. Thus, T&G should not disregard but instead consider similarity as a conceptual basis for their Bayesian categorisation model.

The second problem with T&G's Bayesian approach is that it is incomplete. This is because it only considers the likelihood and leaves out considerations of prior constraints on the inference of categories. For instance, preferences for some over other categories might be important for deciding about cases in which the observed evidence confirms multiple candidate categories equally well.

A final problem with T&G's approach is that the size principle invites worries about undergeneralisation. Undergeneralisation occurs when a to-be-learned word is applied to only a subclass of the items that denote its true meaning. For example, if the word 'dax' meant 'tulip', but a learner applies it only to yellow tulips, then 'dax' is undergeneralised. In Eq. 2.1, undergeneralisation is to be expected because a learner should prefer to generalise towards the smallest possible category that is compatible with the evidence. This is problematic for the empirical adequacy of a categorisation model because we know that category learners, such as children, do not only undergeneralise but often do overgeneralise word meanings to broader categories than would be accurate (cf. Bloom 2002, 36, 158). It is not clear how T&G's model can capture an optimal trade-off between under- and overgeneralisation.

[3] I thank Wolfgang Schwarz for pointing me towards this issue.

2.5 A Geometric Principle of Categorisation

This section uses the conceptual spaces theory to move beyond the size principle. In particular, the argument is that the size principle can be made more precise by considering a conceptual space as a semantic basis for Bayesian inference. I start by outlining Decock et al.'s (2016) argument that the conceptual spaces theory can be used to formulate a geometric principle of indifference as a solution to the different carvings-up problem. I subsequently argue that Decock et al.'s geometric solution to the different carvings-up problem can be generalised towards a solution to the problems with T&G's approach. This generalisation is framed as a geometric size principle that moves beyond T&G's size principle in making it more precise. Though being motivated by a similar idea, the geometric size principle also moves beyond Decock et al.'s approach in that it can account for category learning.

2.5.1 Decock et al.'s Geometric Principle of Indifference

Decock et al. (2016) use a conceptual space to establish a geometric principle of indifference (gPOI), and thereby avoid some of the problems that were encountered by a standard principle of indifference (sPOI). The sPOI states that

> …given a set of mutually exclusive (at most one can be true) and jointly exhaustive (at least one must be true) propositions, and barring countervailing considerations, one ought to invest the same confidence in each of the propositions. Put differently, given a set of propositions of the aforementioned kind, if you lack any reason *not* to treat them evenhandedly, you *should* treat them evenhandedly. (Decock et al. 2016, 55)

A problem for the sPOI is the different carvings-up problem. This is the problem of choosing the right kind of hypotheses space. For example, considering a box with an unknown number of multi-coloured marbles, one can carve up the space of considered possibilities for any outcome in different ways. Probabilities could be distributed over the set of hypotheses 'red' and 'any other colour'. Or, alternatively, one could distribute probabilities over the set of hypotheses, 'red', 'blue' or 'any other colour'. The first option can be expressed as $\mathscr{H} = \{H_1, H_2\}$, where $H_1 =$'red' and $H_2 =$ 'any other colour'. The second option can be expressed as $\mathscr{H} = \{H_1, H_3, H_2\}$, where $H_1 =$ 'red', $H_3 =$ 'blue', and $H_2 =$ 'any other colour'. Which set of hypotheses is considered matters for the distribution of the probability assignments because these are mutually dependent.

The problem is that these different ways to carve up the space are each probabilistically admissible but taken together, they become incoherent. Considering the first hypotheses space, one could assign $Pr(H_1) = .6$ and $Pr(H_2) = .4$. But considering the second hypotheses space one could assign $Pr(H_1) = .25$, $Pr(H_2) = .5$ and $Pr(H_3) = .25$. How can it be rational to assign $Pr(H_1) = .6$ and $Pr(H_1) = .25$ simultaneously? Intuitively, the same probability value should be assigned to H_1 in each case.

Decock et al. solve the different carvings-up problem in two steps. In a first step, they regard the geometry of concepts as primary to the formulation of the degrees-of-belief functions. For this, they use the architecture of concepts as suggested by the conceptual spaces theory (Gärdenfors 2000).

A conceptual space is a geometric similarity space. It is defined as a number of quality dimensions (e.g. height, size, hue, saturation, brightness). Objects can be assigned a value along each dimension, where each value represents the respective perceived quality of the object. In combining those values from each axis of the conceptual space, objects can be represented as vectors in conceptual space. The distances between the vectors express the perceived similarity between them. Roughly, the larger the distance, the less similar the objects are.[4] Regions in conceptual space represent concepts – cognitive categories – and cover areas in conceptual space. The content of a region captures not only information about already observed members of a category, but also information about yet unobserved members. Thus, a region in conceptual space indicates a concept's intensional content, where the intension could be understood as all the possible qualities that a member of the concept can be assigned (cf. Carnap 1988).[5]

On Decock et al.'s account, the basic quality dimensions are important for specifying the gPOI. This is because they are taken as the fundamental attributes (e.g. attributes such as *shape, size, colour*) needed to define the predicates used in the formulation of the hypotheses.

In a second step, Decock et al. use a modification of Carnap's (1980) γ-rule to specify the prior probabilities for any possible carving up. The γ-rule specifies how the probability of an object, o's (e.g. an unobserved colour shade), falling in a region, C_i, can be computed. The $gamma$-rule says that this probability is equal to the size of C_i relative to the size of the conceptual space, CS (which in Carnap's terminology is the attribute space). Where the sentence 'an object o's falling in a region C_i' refers to the content of a hypothesis, $H_{C_i} : \{o \in \text{region}_{C_i}\}$, the rule can be expressed more formally as follows.

[4]To be a bit more precise, the distance function is exponential. Gärdenfors (2000) suggests the Euclidean distance metric for Euclidean space and the Minkowskian distance metric for non-Euclidean space.

[5]Categorisation in conceptual spaces follows a function that maps each point in similarity space onto a unique cell in a Voronoi tessellation. The fundamental categories that serve as candidates when formulating the hypotheses prior to their evaluation thus result from a mechanism of concept acquisition which requires justification itself. For the purpose of this chapter, I mainly disregard the Voronoi tessellation as a concept acquisition mechanism, and consider only the geometric properties of an established conceptual space as substantial for the argument.

Carnap's γ-rule

$$Pr(H_{C_i}) = \frac{size(region_{C_i})}{size(CS)} \tag{2.2}$$

Decock et al. (2016) argue that Eq. 2.2 can solve the different carvings-up problem because it helps to fix the probabilities for each considered hypothesis (e.g. H_{C_i} : $\{o \in \text{region}_{C_i}\}$ versus H_{C_j} : $\{o \in \text{region}_{C_j}\}$) in dependence of the size of the underlying region (e.g. region_{C_i} versus region_{C_j}). Thus, considering two possible hypotheses spaces, it can be exactly decided how the space should be carved up because the concepts used for formulating the relevant propositions are now fixed in their relative sizes.

Consider the following example. Take C_1 to stand for the region in conceptual space representing the category *red*. Take C_2 to stand for the region in conceptual space representing the category *blue*. And take C_3 to stand for the region in conceptual space representing the category *any other colour*. Following Eq. 2.2, the prior probability of a hypothesis pointing towards an object o to lie in C_1 is the area of C_1 divided by the area of the entire conceptual space. Decock et al. (2016) call this prior probability measure α. Then, it does not matter how the space is carved up, that is, whether the partition $\{(o \in C_1), (o \in C_2)\}$ or the partition $\{(o \in C_1), (o \in C_2), (o \in C_3)\}$ is considered. The assignment of a prior degree of belief to $(o \in C_1)$ will be the same, namely α, regardless of whether any other colour is considered in the prediction. With a similar example, Decock et al. show that by fixing probabilities based on the spatial properties of the conceptual space, it is possible to stay neutral on which partition is the right one – the prior degrees of belief stay the same.

Decock et al.'s approach illustrates a further advantage of the conceptual spaces theory. This is that the geometric space can help making the γ-rule more precise. In particular, Decock et al. use the geometric properties of the conceptual space to establish a unique measure of the size of a region – the Lebesgue-measure. Thereby, they can make more explicit the relationship between the assignments of probabilities to candidate hypotheses (i.e. beliefs) in dependence of their semantic contents (i.e. areas in conceptual space). More formally, their modification looks as follows.

Decock et al.'s Lebesgue-specification of the γ-rule

$$\mu^*(C_i) = \frac{\mu(C_i)}{\mu(CS)} \tag{2.3}$$

For Decock et al., the normalised Lebesgue measure, μ^*, represents the prior degree of belief of the agent to classify an unknown object as a member of concept C_i. Thus, Eq. 2.3 makes Eq. 2.2 more precise. This is important because just upon considering the conceptual space, it is possible to semantically interpret the hypotheses. This is because the conceptual space provides the basic predicates without which the hypotheses could not be formulated. The specification is thus possible only via the assumption of the conceptual space. In particular, the $size$ is measured by integrating over the subset of vectors in conceptual space that would be covered by each category predicate, C_i. This way, the gPOI delivers a formally precise solution to the different carvings-up problem, and presents a better alternative to the sPOI.

More generally, Decock et al.'s account is similar in its spirit to the criticism of T&G's approach presented here. Roughly, just like the sPOI lacks a semantic basis for prior degrees-of-belief functions, T&G's size principle lacks a semantic basis for determining the likelihoods in Bayes' Theorem. The common claim is that the conceptual spaces theory offers a way out of the problem by making Bayesian inference more precise, and semantically interpretable.

2.5.2 Going Beyond the Size Principle

To avoid the problems with the size principle, I follow Decock et al.'s approach in two steps. The first step is to define the hypotheses space with the conceptual spaces theory. The second step is to explore in which ways the architecture of a conceptual space can provide additional constraints on categorisation.

For the first step, the conceptual spaces theory suggests that a category should be interpreted as a region in the conceptual space. It can be seen that under this interpretation, the size principle can be compared to Carnap's γ-rule. Equation 2.2 takes the size of the relevant region relative to the size of the entire conceptual space. In contrast, Eq. 2.1, disregarding for a moment the $|n|$-component, reverses this relation. An assignment of total probability (which represents that the observed item belongs to any category) is taken relative to the size of a candidate category. Thus, the size principle reverses the influence of the size of a category on the degree of belief function. Given the structural similarities between Eqs. 2.2, 2.3 and 2.1, also the size principle can be specified with the help of the geometry of concepts. Following Decock et al., the surface area of a region in conceptual space can be measured with the Lebesgue measure, μ (see Eq. 2.3). One way in which this could be expressed is as follows.

$$Pr(e_i|H_{\langle C_i,L\rangle}) = \frac{\mu(e_i \cap C_i)}{\mu(C_i)} \tag{2.4}$$

Equation 2.4 expresses the likelihood of a known item, e_i, given that it is a member of a region, C_i, in conceptual space. The likelihood is a measure of the relative overlap of a piece of evidence, $e_i \in E$, and a candidate region, C_i. This is taken relative to the measure of the region, C_i.[6] The likelihood hence becomes larger the better the relative overlap of the evidence with the candidate concept.

Equation 2.4 moves beyond the size principle by answering the problems that were associated with T&G's approach. It can solve the first problem of disregarding similarity too hastily because in contrast to T&G's argumentation, it does welcome similarity as a basis for probabilistic inference in categorisation. This is positive for avoiding the two facets that come with this problem. For the first facet, the geometric approach can offer at least an explanation sketch of how perceived similarities in varying examples relate to the corresponding probability assignments to possible future exemplars. In a conceptual space, exemplar variability is captured by measuring the average distance between the exemplars. A probability density function can be approximated once sufficiently many exemplars have been observed. In a conceptual space with continuous quality dimensions, the probability of observing an item from a particular region in conceptual space is then simply the area under the probability density function bounded by the region. A relation between the probability and the perceived similarities of the exemplars can thus be established based on the distance measure that a conceptual space is equipped with. In effect, this can give rise to the dynamics between the size of a category and how it determines the likelihood for the corresponding hypothesis (cf. Krumhansl 1978).

By considering similarity as a basis for probability assignments, the geometric approach also helps to find a more transparent semantic interpretation of the size principle. This is based on two changes that come with a geometric approach to the size principle. First, Eq. 2.4 replaces the likelihood with a measure of the relative proportion of areas in conceptual space. This defines the content of the hypotheses more precisely: as mappings from measurable regions in conceptual space (e.g. the concept TULIP) to the labels whose meaning is to be inferred (e.g. 'dax'). Second, Eq. 2.4 lets the term for the evidence, E, re-occur on the right-hand side of the equation. In doing so, the geometric approach helps to make more explicit the relationship between the evidence and a hypothesised category (here measured by their relative spatial overlap). Based on these two changes, Eq. 2.4 is easier to interpret than Eq. 2.1. The evidence is a point in conceptual space and a candidate category is a region in such a space, and their relative overlap can be measured geometrically. More generally, this shows that the geometric approach offers a more transparent way to interpret the likelihood term in the Bayesian inference of categories.

To make this point more clear, consider the following example. Zoey's dad wants to teach her the names of some colour categories. He shows her three particular colour shades, $blue_1$, $blue_2$ and $blue_3$, and he calls each of them 'azure'. For an overview:

$E = e_1 : \langle blue_1, \text{'azure'}\rangle, e_1 : \langle blue_2, \text{'azure'}\rangle e_3 : \langle blue_3, \text{'azure'}\rangle$

[6]This formal solution – to consider the relative overlap of the evidence and the candidate category – has already been suggested in a joint talk held by Peter Brössel and me at a conference in Salzburg, 2015 (Poth and Brössel 2015).

$$\mathscr{H} = H_1 : \langle turquoise, \text{'azure'} \rangle,\ H_2 : \langle light\ blue, \text{'azure'} \rangle,\ H_3 : \langle blue, \text{'azure'} \rangle$$

With the help of the conceptual spaces theory, a hypothesis can now be interpreted in terms of the average spatial distance between those subsets in conceptual space that are indicated by the corresponding regions. For instance, the content of H_1 can be partially expressed in terms of the distance of those vectors in colour space that would be perceived as turquoise colour shades. The other part of this content is the linguistic description itself, 'azure', which is given in the supervised learning environment that is considered here. In light of this interpretation, hypotheses can be evaluated in terms of the intensions of the categories that they point towards, that is, in dependence of their average spatial distances. Following Eq. 2.4, the information about perceived similarities as represented in conceptual space can then be used to assign probabilities to each candidate hypothesis. For instance, the semantic content of the hypothesis that $blue_1$, $blue_2$ and $blue_3$ belong to category *turquoise* is the relative overlap of the distance between $blue_1$, $blue_2$ and $blue_3$ and the average distance of perceivable items within the candidate region in conceptual space that represents the intension of the corresponding term (e.g. 'azure').[7] Then, $Pr(E|H_1)$ would win in the competition because it presents the best overlap with E.

Another way in which the geometric approach moves beyond the size principle is that it can make T&G's Bayesian inference model more complete. This is because whereas T&G yet only specify the likelihoods in the inference of categories, a conceptual space can offer additional constraints to determine the prior probabilities. One constraint is convexity. In the conceptual spaces theory, convexity means that if two items in conceptual space are known to belong to the same category, then any other item that would lie on a straight line between the two examples will be known to be a member of the category, too. Convexity can offer an additional constraint on word-meaning inferences in the context of categorisation tasks. This can be helpful for evaluating contexts in which the evidence is insufficient to decide for a unique way to generalise, e.g. when candidate categories are confirmed equally well by the evidence alone.

The motivation to build a convexity constraint in the architecture of a categorisation model comes from the idea that convex categories should be preferred in meaning inferences because they are easiest to infer. In the conceptual spaces theory, natural properties are considered to be those that are convex. This statement is expressed as "Criterion P" in the conceptual spaces theory (Gärdenfors 2000, 71). The rationale behind Criterion P is that it serves to establish a cognitive distinction between natural and gerrymandered categories. Such a distinction is needed to tackle Goodman's (1972) new riddle of induction: why do we want to prefer inferences towards categories such as *green* or *blue*, as opposed to those

[7]The k-nearest neighbour rule (Russell and Norvig 2002, ch. 18) suggests a similar interpretation. Here, to-be-classified items are grouped based on their distance to an average observation in a vector space, and this can be used to determine a model's graded meaning inferences.

towards *grue* or *bleen*? For Goodman, the answer is that *green* and *blue* are projectible predicates and should thus be preferred over *grue* or *bleen*, which are inductively useless. In the conceptual spaces theory, the conceptually abstract notion of projectibility is made more precise with the more spatially grounded notion of convexity.

To see how convexity can help narrowing down the candidate categories if the evidence alone cannot do so, imagine the following example. Over a period of three days, Zoey's dad shows her a sequence of three flowers. The flowers are of different colours but all have the same shape. On each of these observations, Zoey's dad says 'dax'. Thus, there is $E = e_1 : \langle$ white tulip, 'dax' $\rangle$, $e_2 : \langle$ yellow tulip, 'dax' $\rangle$, $e_3 : \langle$ red tulip, 'dax' $\rangle$. What is more, all observations are made during the afternoon. Given these three pieces of evidence, the following hypotheses seem plausible. $\mathscr{H} = H_1$: 'dax' means *tulip*, H_2 : 'dax' means *flower*, H_3: 'dax' means *tulip in the afternoon*.

Intuitively, H_1 is to be preferred over H_2 because H_1 is more plausible in light of E. Following Eq. 2.4, H_1 is better supported by E than H_2 because E and the candidate category *tulip* achieve a better relative overlap in conceptual space than E and the candidate category $flower$. However, E is insufficient to give H_1 an advantage over H_3 because all observations were in fact made in the afternoon. But intuitively, H_3 also appears less plausible than H_1. Equation 2.4 cannot provide sufficient reason for this because it only considers the role of the evidence in determining the probabilities. Thus, some other criterion must be chosen to evaluate the hypotheses in addition to the available evidence.

The convexity constraint can help here. The category *tulip in the afternoon* is most probably non-convex because it mixes the category *tulip* with the time dimension. Whereas objects that look like tulips are likely to form convex clusters in the shape and colour domains, adding the time dimension would make them become scattered in conceptual space. This is because by adding the constraint *in the afternoon*, objects that would be classified as *tulip* would lose the property of being a tulip during some intervals along the time dimension (e.g. during those intervals that represent the morning and the evening).[8] Under the assumption that *tulip in the afternoon* is non-convex, H_3, as opposed to H_1, points towards a gerrymandered category and should thus be harder to infer. In other words, H_1 should be preferred over H_3 because it points towards a more natural category, even if the evidence confirms both hypotheses equally well.

Formally, this suggestion will be best captured in the prior probabilities. That is, before Zoey is going to observe the next example, her prior degree of belief in the

[8]See also Gärdenfors' (2000, pp. 211) solution to the grue- and bleen problem, in which he argues that *grue* and *bleen* are non-convex because they are defined by the hue and the time dimensions. He argues that in contrast, the categories *blue* and *green* are convex because in the latter cases there is no interference with the time dimension.

hypothesis that ‘dax’ means *tulip* should be higher than her prior degree of belief in the hypothesis that ‘dax’ means *tulip in the afternoon*, just because *tulip* is a more natural concept.[9]

Apart from the specification of the prior probabilities, criterion P could also help with the worry of undergeneralisation. This is because if a category is convex, then it will have to occupy on average a larger proportion of the conceptual space than a category that maximally fits the examples. Though a full reply to this problem cannot be fleshed out here, one option is to consider the Voronoi tessellation (Gärdenfors 2000, pp. 87) to accommodate overgeneralisation. A Voronoi tessellation partitions the conceptual space into clusters of mutually exclusive and exhaustive convex sets of objects. The tessellation process starts from a prototype. At least two prototypes are needed to achieve a tessellation because the cells are established through a connection of the bisectors of hypothetical lines that connect the prototypes for any cluster of items. Upon tessellating the space, any item that is closer in space to a prototype than to any other prototype will be assigned a membership function for the category that is represented by the prototype, i.e. the cell associated with that prototype. With this method, a learner would overgeneralise easily, if the partition is broad enough. Generally, the approach outlined here should not be committed to the Voronoi tessellation. But it illustrates that convexity can indeed help to counter the worry of undergenralisation that comes with the size principle.

To recapitulate, this section has taken Decock et al.’s solution to the different carvings-up problem as starting point to help moving beyond the size principle. The advantage of the resulting geometric approach to categorisation is that it can avoid the problems presented in Sect. 2.4 while keeping in with Shepard’s original motivation to use similarity as a tool to explain the generalisation gradient.

2.5.3 A Worry for a Geometric Principle of Indifference

Decock et al.’s approach has a problem; it cannot explain how information about how a category is commonly used changes the degree of belief in that category being the right meaning candidate. That is, it fails at accounting for pragmatic effects in category learning.[10] For instance, consider two categories, C_a and C_b, represented as regions in conceptual space. Imagine that the size of both corresponding regions is the same, as measured by the Lebesgue integral. That is, $size(C_a)/size(CS) = size(C_b)/size(CS)$. Because according to Carnap’s γ-rule, all that matters for determining the prior degree of belief is the relative size of a category, the prior probabilities for C_a and for C_b must also be the same. Based on Decock et al.’s

[9]There is some evidence for a preference of convex categories in children learning homophones (e.g. Dautriche et al. 2016) and also amongst adults and other word types more generally Jäger (2010).

[10]Thanks to Peter Gärdenfors for pointing out this problem to me.

approach, we obtain that $\mu^*(C_a) = \mu^*(C_b)$ and thus that for any unknown item o, $Pr(o \in C_a) = Pr(o \in C_b)$. But imagine that it is also known that C_a is used often in linguistic communication, whereas C_b is very rarely used. It is not clear how this information could be captured by a measure of the size of either C_a or C_b alone. Thus, prior probabilities alone cannot capture the pragmatic influences on degrees of beliefs about category membership for unknown objects.

More generally, the problem is that Decock et al.'s approach cannot account for category learning. This is because it is limited to a static account of rationality, that is, one in which an agent is rational at a given point in time if and only if her degrees of belief can be represented by a probability function at that point in time. However, learning requires beliefs to change in response to novel evidence. Thus, for a rational-learning account of categorisation, a dynamic account of rationality that considers the evidence is needed. On such an account, an agent is rational if and only if her change in a degree of belief from an earlier to a later time point can be represented by conditionalisation.[11] That is, her degree of belief in H conditional on E prior to E's occurrence, $Pr(H|E)$, must equal her degree of belief in H after having learned E, $P_E(H)$. But to compute $Pr(H|E)$ and consider what happens upon learning E, a rational learner must follow Bayes' Theorem and compute the likelihood.

My suggestion is that information about how frequently a category is used is part of the evidence for category learning, and that such information can be captured in the likelihood term. Then, the geometric size principle would already contain the ingredients to accommodate the pragmatic challenge because it specifies the likelihood term.

A possible response from the geometric size principle could be made by stressing two points. First, the geometric approach specifies the likelihood instead of the prior probabilities. Second, it can also capture the role of multiple examples for category learning. Based on their differences, the geometric size principle is promising in accommodating for pragmatic effects in categorisation.

For the first point, I suggest that Decock et al.'s specification of Carnap's γ-rule and the modified size principle function as complementary elements in a Bayesian model of category learning. Formally, both equations express two different kinds of degrees of beliefs. Whereas the γ-rule replaces the prior probability of observing an unknown member, o, of a candidate category, $Pr(o \in C_i)$, the geometric size principle replaces the likelihood of observing an actual object, e_i, given that it is a member of a particular candidate category, $Pr(e_i|e_i \in C_i)$, where o and e_i can occupy the same point $\langle x, y, z\rangle$ in conceptual space.

For the second point, one could modify Eq. 2.4 to express the idea that the likelihood of observing a sequence of examples for a labelled category becomes proportionally greater the better the relative overlap of the sequence of examples is, on average, with the candidate region in conceptual space.

[11] The best argument for conditionalisation are diachronic Dutch Books (cf. Teller 1973).

$$Pr(E = \langle e_i, \ldots, e_n \rangle | E \in C_i) \propto \left(\frac{\mu^*(\langle e_i, \ldots, e_n \rangle \cap C_i)}{\mu^*(C_i)} \right) \tag{2.5}$$

In intuitive terms, Eq. 2.5 says that the more examples are given, the greater the likelihood for hypotheses that suggest categories that on average overlap well with the examples. This modification captures the role of exemplar variability – the added value from T&G's expansion of Shepard's original model of categorisation. It not only considers the size of a region, but also takes into account the relative locations of category examples, and the frequency at which they are observed.

The modification can help a reply to the pragmatic challenge. If one category, C_a is used more frequently in the learning history than another category, C_b, the likelihood for the former must, overall, be greater than the likelihood for the latter. If C_a and C_b are mutually exclusive regions in conceptual space, the likelihoods equal the relative frequencies at which points in these regions are observed. In a set of 10 examples, $E = \langle e_1, \ldots, e_{10} \rangle$, 7 examples are called 'dax' and overlap with C_a and 3 examples are called 'fep' and overlap with C_b. Thus, 'dax' = 7/10 and 'fep' = 3/10. Thus, $Pr(E, \text{'dax'} | E \in C_a) = .7 > Pr(E, \text{'fep'} | E \in C_b) = .3$. This means that in terms of the likelihood, a rational categoriser should favour a hypothesis that points towards C_a as the more commonly used concept.

This approach could be made more precise by the outlined combination of the conceptual spaces model and probabilistic inference. Inferences as to which category any unknown object belongs to could be expressed by a probability density function that runs over the conceptual space as estimated based on the density of the meaning examples. The more labelled examples for one category as opposed to another are given, the higher the probability density for the corresponding area in conceptual space. For instance, upon observing $\langle e_1, \text{'dax'} \rangle$, the difference in the relative overlap with C_a and with C_b might not be significant. Given multiple examples, however, this difference should increase and make the corresponding prediction for observations of any next item to be called 'dax' more precise. If the next two examples for 'dax', e_2 and e_3, are also in C_a then e_2 and e_3 present evidence that confirms the hypothesis that 'dax' refers to the region C_a relatively more than the hypothesis that 'dax' refers to the region C_b in conceptual space. Thus, given the evidence, the probability density for an area confined to C_a should be greater than the probability density for C_b. Following the original intuition behind T&G's size principle, the relative overlap between E and C_a becomes larger over time because C_a is used more frequently throughout the agent's learning history. Thus, even if C_a and C_b would have the same size, C_a might be confirmed better by the evidence over time than C_b.

2.6 Conclusion

This chapter has outlined two conflicting views on the role of generalisation probabilities in perceptual categorisation. On the one hand, there is Shepard's (1987) view that the probability of generalisation is a derivative of perceptual similarities. On the other hand, there is Tenenbaum and Griffiths' (2001) view that the probability of generalisation governs perceived similarities. I have argued that the theory of conceptual spaces (Gärdenfors 2000, 2014) can be used as a semantic mediator between these conflicting views on the role of generalisation probabilities in perceptual categorisation. The wider implication of this approach is that the notion of similarity is conducive to the psychological plausibility of probabilistic models of categorisation and should therefore not be considered irrelevant in the explanation of the generalisation gradient.

Taken together, I have presented three main reasons to consider a geometric size principle valuable. First, it provides a semantically interpretable basis for Bayesian inference. Second, it can accommodate the intuitions behind both T&G's and Shepard's approaches, that categorisation on average sharpens with an increase in similarity and an increasing number of examples such that generalisation becomes more restrictive. This is positive, for example, because it can explain undergeneralisation. Third, the geometric approach provides additional constraints, such as convexity, that could possibly be built into a Bayesian model of categorisation. This would be advantageous to explain, for example, overgeneralisation in category learning.

Future directions call for a way to connect the outlined model with more objective principles of rationality in categorisation and linguistic communication. Further elaboration is also needed with respect to conditionalisation of degrees of beliefs about category membership in conceptual spaces.

Acknowledgements Thanks to Peter Gärdenfors, Alistair Isaac and Mark Sprevak for helpful comments on an earlier draft of this paper. Thanks to Peter Brössel for helpful discussions along the way.

References

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge: MIT Press.

Carnap, R. (1980). A basic system of inductive logic, part II. *Studies in Inductive Logic and Probability, 2*, 7–155.

Carnap, R. (1988). *Meaning and necessity: A study in semantics and modal logic*. Chicago: University of Chicago Press.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, *31*(1), 607–649.

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: Homophony and the distribution of learning exemplars. *Language Learning and Development*, *12*(3), 231–251.

Decock, L., Douven, I., & Sznajder, M. (2016). A geometric principle of indifference. *Journal of Applied Logic*, *19*, 54–70.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467.

Fairchild, M. D. (2013). *Color appearance models*. Chichester: Wiley.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge: MIT Press.

Gärdenfors, P. (2014). *Conceptual spaces: The geometry of meaning*. Cambridge: MIT Press.

Goodman, N. (1972). *Problems and projects*. Indianapolis: Bobbs-Merrill.

Jäger, G. (2010). Natural color categories are convex sets. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, language and meaning: 17th Amsterdam Colloquium*, Amsterdam, The Netherlands, 16–18 Dec 2009, revised selected papers (pp. 11–20). Berlin/Heidelberg: Springer.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*(5), 445–463.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.

Newton, I. (1687/1934). Principia mathematica. *Newton's Principia, 634*.

Poth, N. L., & Brössel, P. (2015). *A Bayesian answer to the complex-first paradox* (unpublished paper). Salzburg Conference for Young Analytic Philosophy, University of Salzburg.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (International Edition). Englewood Cliffs: Prentice Hall Pearson.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*(2), 125–140.

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy, & J. R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale: Erlbaum.[SK].

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology, 1*(1) 1–17.

Teller, P. (1973). Conditionalization and observation. *Synthese, 26*(2), 218–258.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences, 24*, 629–640.

Tversky, A. (1977) Features of similarity. *Psychological Review, 84*(4), 327.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review, 114*(2), 245.