What Are Natural Concepts?
A Design Perspective

Igor Douven
SND/CNRS/Sorbonne University
igor.douven@sorbonne-universite.fr

Peter Gärdenfors
Cognitive Science, Lund University
peter.gardenfors@lucs.lu.se

Abstract

Conceptual spaces have become an increasingly popular modeling tool in cognitive psychology, artificial intelligence, and philosophy. The core idea of the conceptual spaces approach is that concepts can be represented geometrically, as regions in similarity spaces. While it is generally acknowledged that not every region in such a space represents a *natural* concept—a concept that figures or could plausibly figure in human thinking—it is still an open question what distinguishes those regions that represent natural concepts from those that do not. The central claim of this paper is that natural concepts are represented by the cells of an optimally designed similarity space. To explicate the notion of optimal design, we present a number of general desiderata formulated in terms of parsimony, informativeness, representation, contrast, and learnability. Detailed support for the proposal comes from empirical and computational research on color categorization and from agent-based social simulations.

Keywords: concepts; design; meeting of minds; optimality; similarity spaces.

1 Introduction

Consider the concepts BLUE and GRUE, where something falls under the latter concept if it is green and examined before a future time *t*, or blue and not examined before *t* (Goodman 1954). While BLUE strikes everyone as a perfectly natural concept, GRUE appears gerrymandered, even absurd. But what accounts for this difference in status? This question is at the center of a long-standing philosophical debate, but it is also directly relevant to recent developments in the cognitive sciences. Specifically, it arises in relation to the increasingly popular conceptual spaces framework for representing concepts, which emerged from empirical work in cognitive psychology (Nosofsky 1987; Indow 1988; Gärdenfors 2000) but has in recent years also been used with great success in a range of areas of artificial intelligence as well as in philosophy.[1]

[1] Applications in artificial intelligence research include the formalization of commonsense reasoning (Schockaert & Prade 2013; Derrac & Schockaert 2015), computer vision and robotics (Chella, Frixione, &

Central to the conceptual spaces framework is the idea that concepts can be represented geometrically, as regions in so-called similarity spaces (roughly, metrical spaces where the metric measures similarity; more on this below). At the same time, it is clear that not just any region in a similarity space represents a concept, at least not one that plays or could play a role in our thinking and theorizing—not a *natural* concept, like BLUE, and unlike GRUE. However, this raises the question of what differentiates natural from non-natural concepts, a question that has so far received no fully satisfactory answer in the conceptual spaces literature.

Literature on the universalism/relativism debate suggests some prima-facie plausible answers to this question. The way the question has been asked in that literature is whether supposedly natural concepts like BLUE, WATER, and TIGER are universal, reflecting real divisions in reality that must therefore be shared by all people, or whether their extensions are, at bottom, arbitrary, reflecting mere conventions or cultural biases.

It speaks in favor of the former, "universalist" answer that there seems to be something objectively right about our having separate concepts for horses and cows, or about our having the concept BLUE but not the concept GRUE. On the other hand, relativists are in a much better position to explain why we do not find the same conceptual systems across the globe, and even find ones that deviate quite starkly from our own. Following the "meeting of minds" model of Warglien and Gärdenfors (2013), we propose a middle ground between universalism and relativism that, we believe, can explain both how a conceptual system can be, in an important sense, objectively correct and how one might plausibly expect to find deviations from that system.

According to our approach, a *conceptual system* is an agreement between the members of a community that a particular meaning domain be partitioned in a particular way. A *concept* is then a particular element of such a partitioning, for example a color or a fruit concept. The agreement about a conceptual system need not be explicit but typically emerges out of the successes and failures of the interactions between the members of the community (Warglien & Gärdenfors 2013). Such a "meeting of minds" is universal in the sense that all members of the community must adopt the same conceptual system in order to cooperate and communicate with others. It is also relativist in the sense that a variety of conceptual systems are possible for a particular meaning domain. For instance, the number of color concepts varies extensively between different cultures (see Section 5). Furthermore, experts in an area typically use finer-grained conceptual systems than lay people do. This approach makes concepts socio-cognitive constructs rather than Platonic abstract entities (Gärdenfors 2014).

Given this characterization of conceptual systems, one may ask whether a particular system is optimally designed, where "optimality" is defined by reference to certain broad constraints to which we humans are subject. Our central claim is that natural concepts

Gaglio 1997, 2000, 2003), the computational modeling of linguistic phenomena (Steels 2012; Tang & Lawry 2012; Lawry & Eyre 2014; Lewis & Lawry 2014, 2016), and cognitive architectures (Aisbett & Gibbon 2001; Lieto et al. 2015; Lieto, Chella, & Frixione 2017). In philosophy, conceptual spaces have been used in the study of theory change (Gärdenfors & Zenker 2011, 2013), vagueness (Douven et al. 2013; Douven 2018a), graded membership and truth (Decock & Douven 2014; Douven & Decock 2017; Douven 2018b), the nature of perceptual experience (Brössel 2017), and ethical issues related to technological advances (Peterson 2017). See also Roy (2005a, b) and Steels and Belpaeme (2005) for applications of the conceptual spaces framework in computational linguistics, and Kriegeskorte and Mur (2012), Kriegeskorte and Kievit (2013), and Mur et al. (2013) for applications in neuroscience.

are those represented by an optimally designed conceptual system. Contra relativism, this renders the notion of a natural concept non-arbitrary, while also allowing for the occurrence of (sometimes substantial) differences among conceptual systems used in different cultures. But, contra universalism, the best designed conceptual systems, and hence natural concepts, are not held to reflect some fundamental blueprint of the physical or mental world.

If one adopts a design perspective, a crucial question is what concepts are *for*. There are three main uses of concepts: (i) for categorization; (ii) for communication; and (iii) for reasoning. Firstly, we need to categorize entities in the world in order to choose appropriate actions. For example, we must be able to distinguish edible things from non-edible ones. Secondly, we often communicate about concepts, and not only about single objects or individuals. For example, in teaching contexts, general relations between concepts ("elephants have trunks," "smoking causes cancer") are central. Thirdly, concepts are used in reasoning, as has been studied by philosophers at least since Aristotle's syllogisms. In this paper, we primarily focus on design principles—according to which, for example, concepts should be informative, representative, and learnable—that pertain to the first two uses.

In Section 2, we rehearse the main tenets of the conceptual spaces approach. In Section 3, we present various candidates for cognitive criteria on the design of conceptual systems. Section 4 discusses examples of specific design constraints in the context of the conceptual spaces approach. Section 5 makes the discussion more concrete by illustrating how the criteria and constraints function in the color domain. Section 6, finally, takes a social perspective and shows how the same criteria and constraints are also motivated by the need for concepts to serve purposes of communication and of social interaction more generally.

Before we start, we should note that there is a vast and diverse literature on concepts, with contributions from psychologists (Osherson & Smith 1981, 1982, 1997; Medin 1989; Malt 1994, 1995; Medin, Lynch, & Solomon 2000; Malt, Sloman, & Gennari 2003; Murphy 2004; Malt & Sloman 2007; Carey 2009), linguists (e.g., Lakoff 1973; Kamp & Partee 1995; Jäger 2007, 2010), and philosophers (e.g., Rey 1983, 1985; Aydede & Guzeldere 2005; Margolis & Laurence 2007; Machery 2009; Churchland 2012) contributing. Even within each of these research communities, there is much disagreement on what concepts are and even on their ontological status (i.e., whether they are to be interpreted realistically or only instrumentally, as convenient tools for psychological theorizing). For our own view on concepts, including their ontological status, see Gärdenfors (2000, 2014), where differences and commonalities with much of the just-cited literature are also highlighted.

2 Conceptual spaces

The conceptual spaces approach is an expanding research program based on the idea that concepts can be modeled as regions of similarity spaces. Not every similarity space represents concepts (see, e.g., Johnson 2008, Sect. 6.5), but the ones that do are referred to as "conceptual spaces" (Gärdenfors 2000). Conceptual spaces that have been discussed in the literature include color space (e.g., Shepard 1964; Indow 1988; Bosten et al. 2005; Douven et al. 2017), taste space (Churchland 1986), olfactory space (Castro, Ramanathan, & Chennubhotla 2013), various auditory spaces, as well as shape spaces (Petitot 1989; Gärdenfors 2000; Churchland 2012; Douven 2016; Valentine, Lewis, & Hills 2016), musical spaces (Shepard 1982; Bååth, Lagerstedt, & Gärdenfors 2014; Nussbaum 2015), spaces to represent actions, events, emotions, moral concepts, scientific concepts, epistemic concepts, and many more

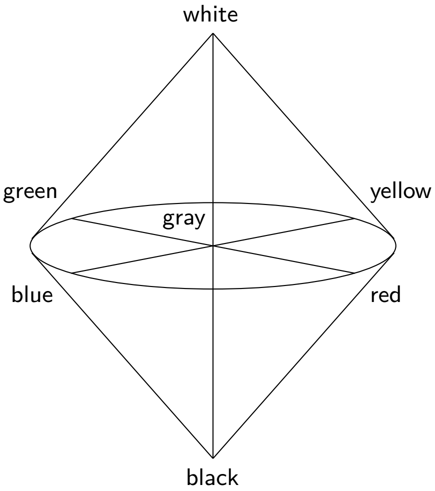

Figure 1: An approximate representation of color space.

(e.g., Gärdenfors 2007; Gärdenfors & Zenker 2011, 2013; Gärdenfors & Warglien 2012; Decock et al. 2014).

As a paradigmatic example, we will use human perceptual color space, of which Figure 1 shows an approximate representation. This space is three-dimensional, with one dimension—the vertical axis—standing for luminance or brightness, which goes from white to black through various shades of gray; the second dimension is the hue circle, which goes through yellow, green, blue, violet, red, and orange, to arrive at yellow again, with these colors gradually blending into each other; and the third dimension is saturation, which is the intensity or depth of the corresponding shade and for any point in color space is given by the shortest distance to the vertical axis.

Similarity spaces are, as theoretical constructions, mathematical entities, specifically, one- or multidimensional structures with a metric defined on them. What makes them *similarity* spaces is that distances in the space are meant to measure dissimilarities between objects. More exactly, the dimensions of these spaces are interpreted as representing fundamental subjective qualities that objects may be perceived to possess to different degrees, so that objects can be mapped onto points in the space in accordance with the degree to which they instantiate each fundamental quality. Distances between such representations of objects are then supposed to measure how similar the objects are to each other, where the similarity is not overall similarity but similarity in the respect—color, taste, shape, and so on—that the space is aimed to model. The metrics most frequently encountered in the literature are the familiar Euclidean distance and the so-called Manhattan distance; the latter simply adds up distances along all the dimensions of a space.[2]

This way of modeling concepts allows them to be context-dependent. For example, in a biological context, an avocado may be classified as a fruit, while in a cooking context is it classified as a vegetable. Another example is that our classification of colors is dependent on the composition of the background light. Such cases of context-sensitivity can be modeled in similarity spaces by assigning different *weights* to the dimensions that are involved in

[2] Metrics based on polar coordinates have also been proposed; see, e.g., Gärdenfors (2014) and Zwarts and Gärdenfors (2016).

the application of the relevant concept or concepts (see Gärdenfors 2000, Sect. 4.2.1; also Krumhansl 1978 and Johannesson 2000).

Neither the shape of a similarity space nor the choice of metric defined on it is arbitrary. Usually, both the structure and the metric of a similarity space are determined on the basis of a large set of similarity ratings, often obtained in an empirical study, which serve as input for one of several related statistical techniques that turn similarities into geometric objects. A frequently used technique is multidimensional scaling (Borg & Groenen 2010); less frequently, other techniques are also used (such as principal component analysis or non-negative matrix factorization; see, e.g., Castro, Ramanathan, & Chennubhotla 2013). By applying such techniques to similarity data, one hopes to achieve not just some arbitrary geometrical representation of those data, but a space that (i) is *low-dimensional*; (ii) has *interpretable* dimensions, that is, dimensions one can make sense of, ideally by associating them with a fundamental attribute; and (iii) has *good fit*, meaning that it faithfully represents the similarity judgments. For a detailed description of the various steps involved in creating a similarity space, see, for instance, Douven (2016).

In principle, similarity judgments for the same set of stimuli could vary widely across subjects, and so two subjects' color similarity spaces (say) could look very different from one another. In practice, there turns out to be large if not perfect agreement in subjects' similarity judgments for a great variety of domains. That is no surprise in view of the fact that, at least for a number of similarity spaces, scientists have been able to relate their structure to specific features of human physiology; for instance, the shape of color space has been explained by reference to the functioning of the rods and cones in our retinas and the early stage of the visual process (see Bosten et al. 2005; Churchland 2012).

3 Design criteria

There is a trend in modern theoretical biology to explain the existence of traits of or processes in organisms, or the workings of biological networks, by reference to good engineering design (Savageau 2001; Alon 2003; Nowak 2006; Poyatos 2012). For example, Alon (2003) points out that the principles of modularity, robustness with regard to component tolerances (roughly, stability under commonly occurring interferences), and use of recurring circuit elements found in many well-designed engineered networks are also found in many biological networks (e.g., groups of interacting cells, or groups of interacting molecules within a cell). The basic principles involved in this kind of theorizing are usually referred to as "design principles," which have been characterized as "patterns of organization that can be specified abstractly, supplying an explanation for a given behavior that occurs across a range of cases in which the organizational pattern is realized" (Green, Levy, & Bechtel 2015, 16). In this paper, we appeal to design principles to characterize natural concepts, specifically as concepts that occur in an optimally designed conceptual system.

The aim of this section is to present and motivate a number of design criteria that a conceptual system should fulfil. We distinguish between *criteria*, which formulate general principles that are independent of particular implementations, and *constraints*, which presume some sort of representational format for a conceptual system. Further on, we look at a number of constraints specifically for similarity spaces, which is the representational format we focus on in this paper.[3]

At the most general level, the approach to be proposed can be described in terms of an engineering task. Suppose we aim at endowing a system with a conceptual scheme, where it is given that the system must succeed in a competition for scarce commodities, and that the system's success depends, inter alia, on its ability to make correct and sufficiently fine-grained classifications. For instance, it should not mistake a poisonous mushroom for an edible one or a foe for a friend, although we are given no specifics about the world the system is to function in (e.g., we are not given a frequency distribution for the different types of mushrooms in that world). The system will have to operate under various general limitations, specifically, that its memory has limited storage capacity, and that its discriminatory powers do not allow it to detect arbitrarily small perceptual differences. Moreover, the system should be able to learn to function on its own in a relatively short time. Finally, the system must be able to communicate with other systems. The overarching goal is to optimize the system's chances of long-term success.

In a first stab at the problem, we translate these givens into a number of very general design objectives. One objective is to provide the system with a rich arsenal of concepts. Another is to minimize strain on the system's memory: the conceptual scheme must be parsimonious. That may also help to partly realize a third objective, which is that the scheme must be easy to learn. Finally, the scheme should be such that it minimizes the risk of classification error.

What follow are core design principles that are meant to pertain to individual conceptual structures and that are in line with the above more general objectives:

Parsimony: The conceptual structure should not overload the system's memory.

Informativeness: The concepts should be informative, meaning that they should jointly offer good and roughly equal coverage of the domain of classification cases.

Representation: The conceptual structure should be such that it allows the system to choose for each concept a prototype that is a good representative of all items falling under the concept.

Contrast: The conceptual structure should be such that prototypes of different concepts can be so chosen that they are easy to tell apart.

Learnability: The conceptual structure should be learnable, ideally from a small number of instances.

Here we briefly comment on these criteria, but they will be further discussed later.

Parsimony is obviously motivated by the system's memory limitations. Informativeness means that the conceptual structure is useful for selecting which actions to perform. For example, if we are looking for a car that functions well under tough mountain conditions, we are not helped by categorizations that build on the color or the CO_2 emission of cars.

[3] To some extent, the distinction between design criteria and design constraints is a parallel to Marr's (1982) distinction between the computational and the implementational level.

Note that there is a tension between these two criteria: the fewer concepts there are in a conceptual structure, the better Parsimony is fulfilled, but the less informative the concepts will be. To balance the two criteria, the connection between concepts and actions must be considered. The set of concepts should be rich enough to allow a system to select the right actions. For example, if you do not know the distinction between black and red elder bushes, then you do not know which flowers to pick to make elderflower cordial. On the other hand, having fine conceptual distinctions may become a burden for your memory. If you are not a professional carpenter, you may, for example, not need to distinguish between curved-claw hammers, ball-peen hammers, drywall hammers, tack hammers, lineman's hammers, and so on. Depending on the environment the system is going to inhabit—which at design time is unknown, as mentioned—and on which more specific goals it aims to pursue (e.g., choosing to become a carpenter), it may have reason to refine its conceptual scheme later on, perhaps in a way that will allow it to make distinctions finer than the ones the initial scheme allows it to make. Representation and Contrast will aid with memorization and with avoiding classification errors. As will be seen later on, these criteria may pull in different directions as well. Learnability, finally, is required since varying environments preclude that all relevant concepts are initially provided.

Further motivation for these criteria will come from considering how they operate in concrete contexts, for instance in structuring the color domain (Sect. 5). Some of the criteria will also be seen to flow from a very general desideratum on concepts as tools for communication (Sect. 6). We are not claiming that the above list of criteria is exhaustive, and we are open to the idea that the notion of an optimally designed conceptual system can be, and may even need to be, explicated in terms of additional design criteria.

4 Concepts as the outcomes of constrained optimization

4.1 Optimal design of vowel space

To introduce the idea of applying design thinking in the context of similarity spaces, we start with a brief discussion of precisely such an application—even if it was not advertized as design thinking—to wit, the explanation of vowel systems as proposed in Liljencrants and Lindblom (1972).

These authors point out that the vowels the human vocal tract is able to produce can be represented in a three-dimensional similarity space (a vowel space), with the dimensions representing the three lowest formant frequencies. While the vocal tract can in principle produce an indefinite number of different vowels, only a very limited number of those possibilities are instantiated in human speech (see also Jakobson 1968). Liljencrants and Lindblom set out to answer the question of why that is so.

For this application, Contrast is a particularly important criterion, an obvious rationale for which is that it minimizes the risk of confusing vowels and thus of misunderstandings in spoken communication. To explain how the notion of perceptual contrast is to be understood in spatial terms, Liljencrants and Lindblom offer a useful analogy from physics: two particles with equal electric charge that can move freely in a container will reach an equilibrium if and only if the distance between them is maximal within the confines of the container. In Liljencrants and Lindblom's proposal, we are to think of vowel space as the container, and the vowels as the particles that seek to maximize their mutual distance.

Evidence for the proposal comes from a computer program that calculates sets of coordinates in vowel space which yield maximum perceptual contrast among n vowels, for $n \in \{3, \dots, 12\}$. The computational results are excellent for the $n = 3$ to $n = 6$ cases in that the computer models predict with great accuracy the vowels found in spoken languages with the corresponding numbers of vowels. The larger vowel systems found by the computer contain more errors. Still, Liljencrants and Lindblom's results are impressive enough to warrant their conclusion that the principle of maximizing perceptual contrast plays a pivotal role in how the inventories of speech sounds of languages are selected, even if—as they remark—the errors for the larger systems are evidence that other factors, not implemented in their model, play a role, too. They (plausibly) speculate that among such other factors may be various articulatory variables, most notably ease of articulation and co-articulability (p. 854). As they put it: "a vowel system which has been optimized with respect to communicative efficiency consists of vowels that are not only 'easy to hear' but also 'easy to say' " (p. 856).

The approach taken by Liljencrants and Lindblom can be applied to the question of how to best furnish similarity spaces with concepts, but—as will be seen below—then consideration should be given to all the design criteria presented in Section 3.

4.2 Design constraints on conceptual spaces

4.2.1 Convexity

We noted that, in the conceptual spaces framework, concepts are regions in similarity spaces, but also that not every region in a similarity space stands for a concept, or at least not one that might ordinarily figure in our thinking or that we might care to name in our language. Thus arose the question of what distinguishes natural concepts from non-natural ones.

As a constraint on the conceptual systems that are based on similarity spaces, the following has been proposed (Gärdenfors 2000, 71, calls the constraint "Criterion P"):

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

That a region is convex means that, for any two points in the region, the line segment between those points lies in its entirety in the region as well. Gärdenfors (2000, 70; 2014, Sect. 7.2) points at important empirical support for Convexity from color-naming studies (notably, Sivik & Taft 1994), which show that what we commonly regard as natural color concepts—blue, green, red, and so on—form convex regions in color space.[4]

Most importantly for our present concerns, Gärdenfors (2000) defends Convexity as "a principle of cognitive economy; handling convex sets puts less strain on learning, on your memory, and on your processing capacities than working with arbitrarily shaped regions" (p. 70). While there is no explicit reference to overarching design criteria here, it is manifest that Convexity was motivated by the same intuitions that underlie Parismony and Representation. Moreover, Gärdenfors points out that if Convexity holds, then by learning of a small set S of items that they fall under a given concept C, one automatically learns of

[4] Recent empirical support not concerning the color domain comes from studies reported in Douven (2016), which investigated the concepts bowl and vase in a shape space. The aim of that paper was to determine degrees of bowlhood and vasehood for a great number of shapes. It was verified that if a degree of membership greater than .5 is chosen as a criterion for belonging to a concept, then bowl and vase both come out as convex in the relevant similarity space.

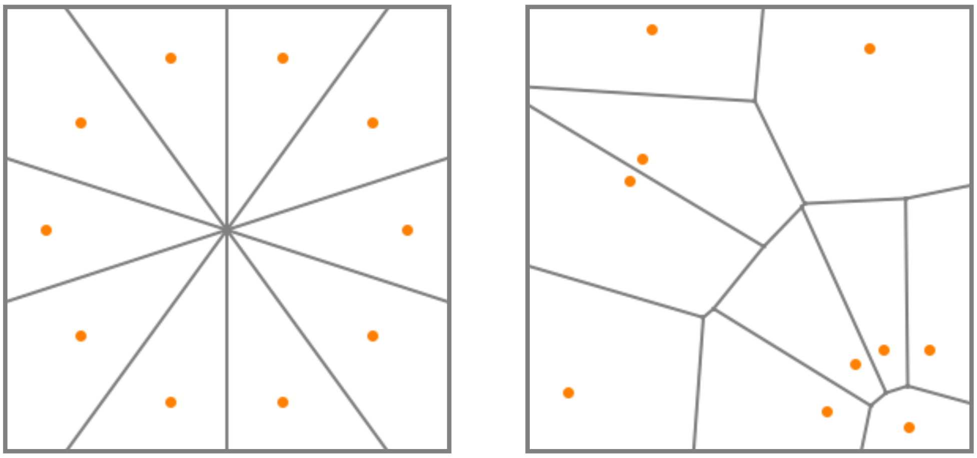

Figure 2: Two Voronoi tessellations of a two-dimensional space. The points represent the prototypes and the lines show the borders between the categories.

all items represented by points in the convex hull of S that they fall under C as well. Thus Convexity is supported by Learnability, too.[5]

But although plausible, the question is whether Convexity is *enough* to single out the natural concepts. Gärdenfors notes that he "only view[s] the criterion as a *necessary* but perhaps not sufficient condition on a natural property" (*ibid.*). To see why Convexity is in fact *not* enough to differentiate natural from non-natural concepts, note that every plane intersecting the color space shown in Figure 1 divides that space into two convex regions. The property of convexity is also preserved under set intersection (see, e.g., Douven et al. 2013, 147). Thus, we can randomly pick any number of planes intersecting color space and be guaranteed to end up with a division of that space into convex regions. What holds for color space holds for almost any similarity space: there are countless ways of partitioning the space into convex regions that yield concepts, few of which will count as natural in any intuitive sense.

4.2.2 Prototypes to the rescue?

Gärdenfors' own brand of the conceptual spaces approach relies heavily on prototype theory, and it might be thought that prototypes are exactly what we need to answer the question of how similarity spaces get equipped with concepts. Specifically, Gärdenfors combines prototype theory with the mathematical technique of Voronoi tessellations. Given a space S and a set $P = \{p_1, \dots, p_n\}$ of points in S, the Voronoi tessellation of S generated by P is the set of cells $\{c_1, \dots, c_n\}$ such that each c_i contains all and only points in S that are at least as close to p_i as they are to p_j, for all $j \neq i$, where closeness is measured by the metric associated with S (see Figure 2).

Thus, imagine that we have the prototypes of red, blue, green, and so on, located in color space. Then the Voronoi tessellation generated by those points divides color space into a number of regions such that the shades closer to the red prototype than to the other

[5] Convexity is also immediately helpful in showing why grue is not a natural concept: it does not correspond to a convex region in any known similarity space (Gärdenfors 1990).

prototypes are grouped together, the shades closer to the BLUE prototype than to the other prototypes are grouped together, and so on. This sounds like we might have identified the natural color concepts, in particular since it is provably the case that the cells of a Voronoi tessellation defined on a Euclidean space—any such Voronoi tessellation—are all convex (Okabe et al. 2000, 58) so that CONVEXITY is automatically satisfied.

For a Voronoi tessellation, PARSIMONY is more generally satisfied in that an individual only has to remember the locations of the prototypes to be able to construct the tessellation, from which she can retrieve the concept under which any given item falls in the space. The degree to which INFORMATIVENESS is satisfied depends on the number of color categories in the system and on how "evenly" they cover the space. While the left tessellation in Figure 2 would appear to do well on INFORMATIVENESS, and also on REPRESENTATION and CONTRAST, satisfaction of these criteria is by no means guaranteed, as is apparent from the right tessellation in the same figure. In that tessellation, quite a number of prototypes are located relatively close to each other, thereby jeopardizing CONTRAST. It would also be false to claim that all parts of the space are roughly equally covered in that tessellation: this conceptual system would allow us to make fine-grained distinctions in some parts of the space (especially the lower right part) but only coarse-grained distinctions in other parts, so it does poorly on INFORMATIVENESS. Note, moreover, that many of the prototypes lie much closer to various other prototypes than to some points in the concepts of which they are the prototypes. As a result, this tessellation scores poorly also on the count of REPRESENTATION.[6]

Even if Voronoi tessellations satisfied all our design criteria, however, a central question would remain, namely: Where do the prototypes come from? It is important to note that prototypes *derive* from categories. In particular, a prototype is supposed to be the *best representative*, or *most typical instance*, of a concept (Rosch 1973, 330)—which presupposes that the concept is already in place.[7] See also Gärdenfors (2000, Sect. 4.5), where it is proposed that the prototype is calculated as the mean of all the exemplars of a category that have been encountered. In itself, this simple rule would explain how it is possible to learn a category quickly, given that the mean is defined as soon as the first exemplar is observed. The rule would thus help achieve LEARNABILITY. The problem is that the rule presupposes that the learning process is *supervised*: the system must be provided with the correct categorizations of the exemplars to begin with. So, on this proposal, too, we arrive at the conclusion that categories come first, prototypes second.[8]

[6] It is to be recalled that we are considering the role of these criteria *at design time*, when specific information about the world is still unavailable. Depending on what the world looks like, and specifically on how objects in the world would be distributed in the space, a person or artificial system starting out with a conceptual scheme represented by the left tessellation might reasonably end up with one represented by the right tessellation.

[7] Douven et al. (2013) argue that, for many concepts, there is actually more than one most typical instance, which they use to show how the conceptual spaces framework can accommodate the fact that many concepts are vague. For a different approach to dealing with vagueness in the context of conceptual spaces, see Lawry and Tang (2009). In this paper, we leave the issue of vagueness aside. See Douven (2018a) for a design approach to vagueness.

[8] Another response would be to claim that prototypes are somehow hard-wired in the brain. It has in effect been suggested that the prototypes of BLACK, WHITE, BLUE, GREEN, RED, and YELLOW can be identified with particular neuronal responses in our processing of colors (De Valois, Abramov, & Jacobs 1966; Kay & McDaniel 1978; Kay, Berlin, & Merrifield 1991). But later research cast doubt on this suggestion (De Valois & De Valois 1993), and eventually it was rejected (Abramov & Gordon 1994; Abramov 1997; De Valois, De Valois, & Mahon 2000).

4.2.3 Well-formedness

Convexity is motivated by design thinking: if one had to design a conceptual architecture for a similarity space, one would want it to yield convex concepts, for reasons of cognitive economy and learnability—reasons that derive their relevance from the fact that the human cognitive system is limited in important ways. While, as we saw, Convexity is unable to single out the natural concepts, at least on its own, there are other constraints that can be invoked to flesh out the notion of an optimally designed conceptual system. The following design constraint appears particularly promising:

Well-formedness: The concepts should be "well-formed" in that the items falling under any one of them are maximally similar to each other and maximally dissimilar to the items falling under the other concepts represented in the same space.

Well-formedness can be thought of as flowing directly from Parsimony and Informativeness (Regier, Kay, & Khetarpal 2007; Regier, Kemp, & Kay 2015) and as being motivated by the same considerations of constrained optimization that underlie our design criteria generally: we will be less prone to misclassify two items falling under the same concept as falling under different concepts if these items are always very similar to each other, and we will also be less prone to misclassify two items falling under different concepts as falling under the same concept if these items are always very dissimilar.

Though not put quite in this way, and not usually designated as a design constraint, Well-formedness plays a central role in the literature on unsupervised learning, specifically on clustering algorithms. For instance, many of the best known clustering algorithms, such as k-means clustering and its variants (e.g., Partitioning Around Medoids and neural network algorithms like self-organizing maps and various adaptive resonance theory networks; see Kaufman & Rousseeuw 1990 and Du 2010 for useful overviews), aim at finding clusters in data such that the clustering as a whole simultaneously maximizes intra-cluster similarity and inter-cluster dissimilarity.

We have taken the label "Well-formedness" from Regier, Kay, and Khetarpal (2007), who have also given a formalization of the constraint. We state the formalization here, because it will play a role further on in the paper. Let variables x and y range over possible objects representable in similarity space S, and let P be a categorization of all possible such objects, meaning that $P(x)$ assigns x to one of a number of mutually exclusive and jointly exhaustive regions of S. Furthermore, let sim be the similarity relation defined on S. Then Regier et al. define the "within-similarity" of P as

$$S(P) \quad := \quad \textstyle\sum_{x,y:\, P(x)=P(y)} \mathrm{sim}(x, y),$$

where $0 \leqslant \mathrm{sim}(x, y) \leqslant 1$. They further define the "across-dissimilarity" of P as

$$D(P) \quad := \quad \textstyle\sum_{x,y:\, P(x)\neq P(y)} (1 - \mathrm{sim}(x, y)).$$

The well-formedness $W(P)$ of a categorization P is then defined as the sum of $S(P)$ and $D(P)$. Although Regier, Kay, and Khetarpal propose these definitions in the context of color categorization (see Section 5), the definitions apply generally to similarity spaces, thus providing a quantitative version of Well-formedness. On this version, concepts should be such that their combination—the category system of the space in which they live—maximizes W.

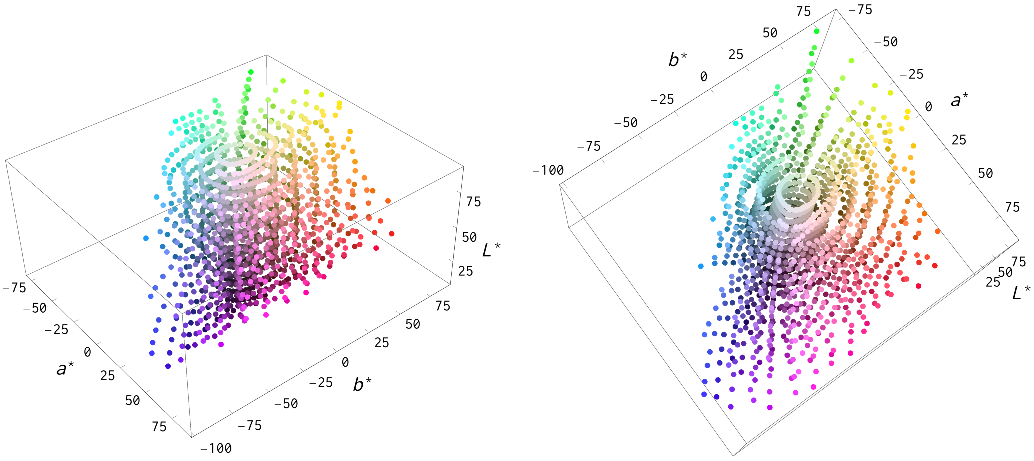

Figure 3: Different views on CIELAB space; the shape of the space is shown by locating in it the set of 1625 Munsell chips that are available from the website of the Munsell Laboratory at the Rochester Institute of Technology.

4.2.4 Anticipated objection

To end this section, we address what may easily appear as a major problem for our proposal, to wit, that our design principles—both the criteria from Section 3 and the constraints just discussed—are *structural* constraints. The concern is that such constraints may not be enough to guarantee uniqueness, in that there may be *many* partitions of a given similarity space that satisfy the constraints to the same maximal extent. Look again at perceptual color space, as shown in Figure 1. It would seem that the design principles considered so far cannot possibly succeed in fixing a unique conceptualization of that space, given that any rotation along the luminance axis is structure-preserving in the relevant sense: any nontrivial rotation along that axis of a structure that satifies our design principles will itself satisfy those principles, although it will yield different color concepts.

Here, it is important to recall that Figure 1 shows an *approximate* representation of color space. In reality, color space is spindle-like indeed, but not nearly as symmetric as is suggested by Figure 1. See Figure 3, which shows the so-called CIELAB space, one of the two main color similarity spaces.[9] The asymmetries in the space are impossible to miss. As a result, the concern about rotational symmetry that would arise for perceptual color space were the representation in Figure 1 exact, does not arise for the actual perceptual color space.

Admittedly, there is no way to be sure that every similarity space will be sufficiently asymmetric to escape the aforementioned problem. On the other hand, there may be additional design constraints that are more directly connected with specific domains, and it is not a priori that such constraints could not be of a non-structural kind. For example, even if the space shown in Figure 1 were an entirely accurate representation of human per-

[9] The other one is CIELUV space, which looks very similar. It is generally assumed that CIELAB space is a more accurate representation of judged similarities between color stimuli when the colors are perceived on paper or on cloth, whereas CIELUV space represents such judgments more accurately when the colors are perceived on-screen; see Malacara (2002, 86–90) or Fairchild (2013, Ch. 10) for a useful discussion of the differences between these spaces.

ceptual space, the world we inhabit might be such that we need to be able to make more fine-grained distinctions in some parts of the space than in others. And if we rotate a partition that achieves that aim, the resulting partition will likely *fail* to achieve it.

Also, while this paper is largely motivated by the thought that not any carving-up of a conceptual space that satisfies Convexity yields natural concepts, and while we agree with Lewis (1983) that natural concepts are *sparse*, it would probably be too much to require absolute uniqueness in all cases. We are understanding the notion of an optimally designed conceptual space as one that does, on balance, best on our design principles. But it cannot be ruled out that, at least for some similarity spaces, there may be more than one way of striking the best balance, and thus it cannot be ruled out that, for those spaces, there is more than one optimal design. Moreover, even if there is a unique optimal design, there may be designs that are so close to being optimal that looking for a still better one may not be worth the effort. Indeed, this is how a constrained optimization approach is compatible with the finding of different conceptualizations across cultures (see below).

5 Design in the color domain

In the previous section, it was suggested that the kind of design thinking embodied in the criteria presented in Section 3 and underlying Convexity and Well-formedness may help us narrow down a conceptual architecture for a similarity space, and thus give content to our central claim about natural concepts. No evidence has been given that the criteria and constraints introduced so far are indeed operative in the process of conceptualization. Here, we look in some detail at color research that not only yields empirical support for the idea that what we think of as natural color concepts correspond to regions in an optimally partitioned perceptual color space, but also supports the hypothesis that specific design principles are actually at work in dividing up color space into concepts.

Berlin, Kay, and their collaborators (Berlin & Kay 1969/1999; Cook, Kay, & Regier 2005) had gathered color-naming data from a great many languages. These data revealed striking universal tendencies in color lexicons as well as noteworthy deviations from those tendencies. Thereby the data put some pressure on both main accounts of color categorization: the universal tendencies were difficult to explain from the relativist standpoint, while deviations from those tendencies were difficult to explain from the universalist standpoint.

Jameson and D'Andrade (1997) put forward an interesting suggestion for explaining both the universal tendencies and the deviations, effectively carving out a sort of middle ground between universalism and relativism. Their suggestion was that both the regularities and the deviations might be due to an interaction between a cognitively motivated preference for informative lexicons and the irregularities in perceptual color space that can be seen in Figure 3 as bumps and depressions. As different lexicons might achieve roughly the same high level of informativeness, Jameson and D'Andrade's explanation leaves some room for cultural influences in categorization, even though the preference for informative naming systems as well as the irregularities in color space are culture-independent, the first being anchored in the efficiency of our cognitive makeup and the latter in our perceptual apparatus.

This suggestion made by Jameson and D'Andrade was put to the test in computational work by Regier, Kay, and Khetarpal (2007), which proposed Well-formedness, presented

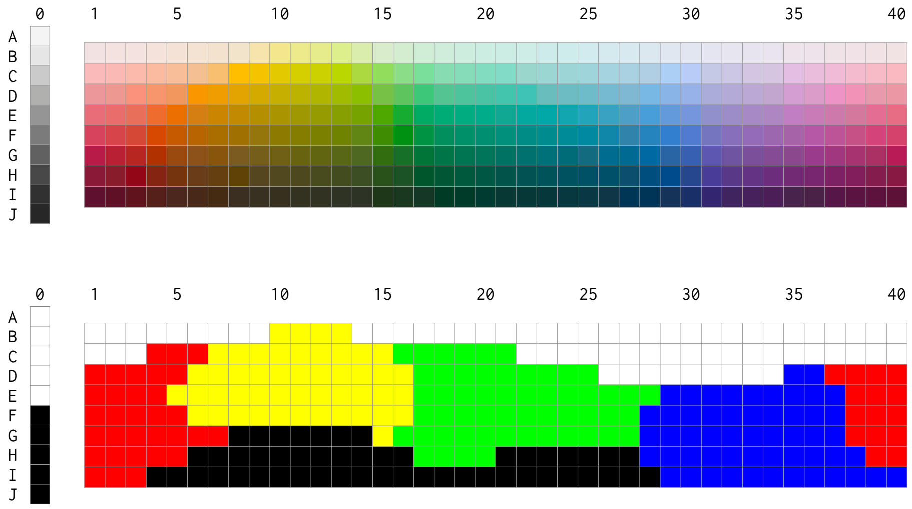

Figure 4: Munsell chips used in Berlin and Kay's (1969/1999) and related studies (top row), and the W-maximal partition of those chips into 6 cells (bottom row).

in the previous section, as capturing the gist of Jameson and D'Andrade's notion of informativeness. Formalized in the way also discussed previously—so in terms of the functions S and D and, ultimately, W—they implemented Well-formedness in computer simulations, and they compared the output from those simulations with the above-mentioned color-naming data.

Regier and coauthors applied the constraint to the 330 so-called Munsell chips that Berlin and Kay used in their studies on color naming. These chips, shown in the top chart of Figure 4, consist of 320 chromatic chips and ten achromatic chips ranging from black to white. The columns of the chart represent equally spaced Munsell hues, and the rows represent levels of luminance; the chromatic chips are all at the maximum saturation available for their hue-value combination (meaning that they are all from the surface of the color space). Specifically, Regier and coauthors' input data consisted of the CIELAB coordinates of the aforementioned Munsell chips. Accordingly, they defined similarity in terms of CIELAB distance, dist, more exactly as an exponentially decaying function of that distance: $\text{sim}(x, y) = \exp(-0.001 \times \text{dist}(x, y)^2)$.

Regier and coauthors used computer simulations to find W-maximal partitions (with 3, 4, 5, and 6 cells) of the Munsell chips; the bottom row of Figure 4 gives the result they obtained for the 6-cell partition. Regier and coauthors showed that these partitions strikingly resemble how a great variety of natural languages categorize the 330 Munsell chips. A further important finding was that partitions which look rather different from the W-maximal partition can still have a W-value very close to the optimum.[10]

Regier and coauthors thereby obtained an explanation of the universal tendencies in color naming found in Berlin and Kay (1969/1999) and Cook, Kay, and Regier (2005) that

[10] Douven (2017) shows that applying Regier, Kay, and Khetarpal's algorithm to the Munsell chips as specified by their CIELUV coordinates leads to even better results.

does without metaphysical heavy lifting by invoking no more than CIELAB distances (so, perceptual similarities) in conjunction with a very general design constraint (viz., Well-formedness) that derives its justification from some well-established bottlenecks in our cognitive makeup. Just as importantly, the computational results suggest a straightforward explanation of the deviations from the universal tendencies that were found in the same color-naming studies: it is easy to imagine how the close competitors to the optimal partition that Regier and coauthors showed to exist may serve our practical interests in ways that are not appreciably worse than those of the optimal partition, so that we will have no incentive to continue our search for a better formed partition once we have one that is "well-formed enough." This could, in turn, be understood as an instance of satisficing behavior in the sense of Simon (1972).

Given that Well-formedness is the only design constraint at play in Regier and coauthors' result, one wonders how it relates to the other design constraint we discussed—Convexity—or to the more general design criteria from Section 2. Does it subsume these, or might a role still remain for these criteria, or for Convexity?

What can be learnt from placing the 330 Munsell chips in CIELAB space is that the partitions for the 3-, 4-, 5-, and 6-cell cases all yield concepts that satisfy Convexity; see Figure 5 for the 6-cell partition. Informativeness is also fulfilled since the cells divide the space into roughly evenly large regions, as is clear from inspecting Figures 4 and 5.[11] And the convex hulls shown in Figure 5 suggest that it should be easy to satisfy Representation and Contrast as well. For instance, we could pick, for each hull, a point centrally located in that hull, and those points would be at quite some distance from one another (relatively speaking). As briefly mentioned in Section 3, however, Representation and Contrast may pull in different directions, and in the case of color space one could imagine putting more emphasis on the second than on the first criterion and choosing prototypes closer to the surface of the space. In reality, that is where color prototypes tend to be located (Berlin & Kay 1969/1999).

With respect to Parsimony, the situation is different. While Regier and coauthors' computer simulations yield conceptual architectures with low numbers of cells, these numbers are put in by hand: Regier et al.'s algorithm, like most clustering algorithms, requires the number of cells to be pre-specified.[12]

We thus see that, in the color domain, Well-formedness is sufficient to realize a fair number of the design criteria discussed in Section 3 and also Convexity. There is no guar-

[11] Or from looking at the numbers of chips assigned to the various clusters. For the 6-cell partition these are: 72 in white, 55 in black, 54 in blue, 54 in green, 47 in red, and 48 in yellow. Equality of coverage can in fact easily be formalized in terms of the Kullback–Leibler distance from a completely flat distribution (Kullback & Leibler 1951). For an n-cell partition P of the 330 Munsell chips, this equals $\sum_i |P_i|/330 \times \ln((|P_i|/330)/(1/n))$, where $|P_i|$ is the cardinality of the i-th cell. For example, for Regier, Kay, and Khetarpal's W-maximal 6-cell partition, the Kullback–Leibler distance from a completely flat 6-cell partition equals 0.01. An alternative here would be the Gini coefficient (Gini 1921), which is best known as a measure of income inequality, but has broader application. For the 6-cell partition, the Gini coefficient is 0.07, which by conventional standards counts as very low (indicating a highly equal distribution).

[12] There are ways to determine an optimal number of clusters, but these are not guaranteed to work, and the present case is one in which they fail; see Jraissati and Douven (2017). This result, as well as other results related to Regier and coauthors' work, might have come out differently if instead of a subset of 330 Munsell chips the full set of those chips (see Figure 3) had been used in the clustering procedure. On the other hand, naming data for that set have only recently become available, and then only for the English language; see Jraissati and Douven (2018).

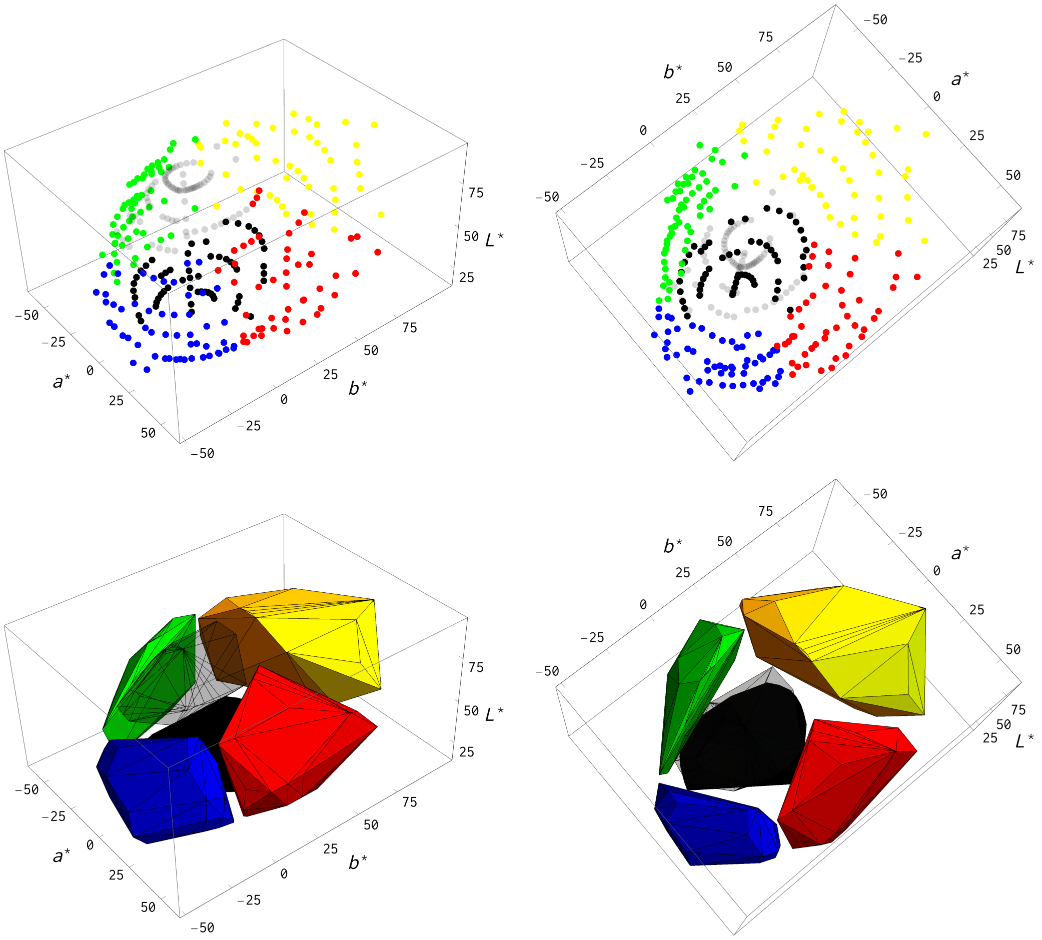

Figure 5: The 330 Munsell chips placed into CIELAB space and colored according to Regier, Kay, and Khetarpal's (2007) result for the 6-cell partition (top row), and the convex hulls of the chips for the various cells (bottom row). (The chips in the white category, as well as the hull of those chips, are shown in light gray.)

antee that this finding generalizes to other domains. Still, Regier et al.'s work is important evidence that design criteria that were presented purely in the abstract have actually been operative in shaping some of our concepts.

6 Design principles from a communication perspective

So far, our design principles have been discussed and motivated by the use of concepts for solving problems of categorization, which includes their use as tools for selecting appropriate actions. In this section, we turn to concepts as tools for communication. This involves taking a social, rather than an individualistic, perspective on the design of conceptual spaces.

A fundamental criterion for a language is that different users mean, more or less, the same thing when they use a word. When learning to speak a language—your mother tongue or a foreign language—a "meeting of minds" must be achieved so that the concepts of a speaker are aligned, via the words of the language, to the concepts of other speakers. This leads us to the following desideratum:

Correlation of meanings: There exists a mapping between the concepts of the speakers of a language so that the meanings of words are correlated between the speakers.

An example of how such a correlation can be accomplished is presented by Jäger and van Rooij (2007), who use computer simulations to show how semantic fixed points (in the form of Nash equilibria) can represent a meeting of minds. They refer to the domain they choose—a circular disk—as "the color space," but there is nothing in the process that depends on relations to colors; in particular, there is no relation between Jäger and van Rooij's "color space" and the CIELAB space previously discussed. The problem they examine is how a common meaning for "color" terms can develop in a communication game. In their example, there are only two players: s (sender) and r (receiver). Jäger and van Rooij assume that the two players have a common space C for "color." There is a fixed and finite set of n messages ("words") that s can convey to r.

The communication game unfolds as follows: Nature chooses some point in the color space, according to some probability distribution over C. The sender s knows the choice of nature, but the receiver r does not. Then s is allowed to send one of the messages to r. In response, r picks a point in the color space. In the game, s and r maximize their rewards if they maximize the similarity (minimize the distance) between nature's choice and r's choice of point. The sender can choose a decomposition S of C in n subsets, assigning to each subset a unique message. For each "color word" sent, there is a prototypical point in the region corresponding to the point that is r's best response. There are thus n prototype points, corresponding to the typical meanings assigned by r to each of the n possible messages from s.

Following the standard definition in game theory, a Nash equilibrium of the game is a pair (S, R), where S is the sender's partitioning (into n subsets) of C, and R is the responder's n-tuple of prototype points of C, such that both are a best response to each other. Jäger and van Rooij (2007) show how to compute the best response functions for each player. The central result of their paper can be restated by saying that if the color space is convex and compact and the similarity function is continuous, then there exists a Nash equilibrium, and it corresponds to a Voronoi tessellation of the color space that is common to s and r.[13] Besides, each prototype point has the property that it minimizes the average distance to all the points in the cell in which it lies. Hence, their solution is already guaranteed to satisfy both Convexity and Representation. Furthermore, the visual presentation of the solution makes it easy to see that it also satisfies Informativeness and Contrast.[14]

In a theoretical analysis, Warglien and Gärdenfors (2013) have generalized Jäger and van Rooij's result by showing how some topological and geometric properties of mental representations make meetings of minds possible. While Warglien and Gärdenfors already assume that concepts can be represented as convex regions of conceptual spaces, they assume neither that the spaces of the communicating individuals are identical nor that they partition the spaces in the same way. Implicitly, their analysis builds on the assumption that language should preserve the nearness relations among points in conceptual spaces, which can be thought of as another design constraint.

[13] That a space is compact means, roughly, that it contains its own borders.

[14] It also satisfies Parsimony, but that is again built in.

It is to be noted that neither Jäger and van Rooij's simulations nor Warglien and Gärdenfors' analytical results guarantee that there is a unique best partitioning of any given similarity space, or even that partitionings achieving Correlation of meanings are sparse, where—to repeat Lewis' point—sparsity is a requirement to maintain that the concepts resulting from the partitioning or partitionings are *natural* ones.

For instance, even though the shapes of the tessellations of the color space C that Jäger and van Rooij arrive at are highly constrained, there still exist infinitely many equilibrium solutions. In particular, since C is symmetrical, all rotations of a solution will also be solutions. But this is a situation we encountered before, when we remarked that every rotation of an optimally designed partitioning of the color spindle shown in Figure 1 would also be optimally designed. It made an important difference, we saw, to consider real perceptual color space rather than the spindle, which only approximates that space. It is thus reasonable to ask whether Jäger and van Rooij's simulations might have resulted in a unique or near-unique solution had they used CIELAB space instead of the perfectly round disk they call "color space."

The answer is that this would probably have made all the difference, as is suggested by the work of Regier, Kemp, and Kay (2015). These authors report the results from a signalling game very similar to that played by the sender and receiver in Jäger and van Rooij's computer simulations, with the receiver also trying to reconstruct the sender's mental representations of specific color shades. The main difference is that in Regier, Kemp, and Kay's game, the receiver's representation error is based on the shape of CIELAB space. Regier and colleagues then show that those partitionings of that space which minimize expected representation error closely resemble ones lexicalized by the languages in the World Color Survey (Cook, Kay, & Regier 2005). In addition to this, they show that, conversely, the partitionings lexicalized by those languages all incur relatively low expected representation errors.

Further support for the role that design criteria play in structuring similarity spaces, according to our account, comes from a number of experimental results concerning language transmission. In a large-scale laboratory experiment, Xu, Dowman, and Griffiths (2013) showed subjects examples of how colors of Munsell chips were named, and the subjects then classified other colors on the basis of the examples. These subjects' responses were used to generate examples for the subjects of the next "generation" of learners. This process continued for thirteen generations. The results reveal that color classifications converge quickly toward color systems similar to those found across human languages. This is a strong indication that these systems satisfy Learnability.[15]

In addition to this, Xu and colleagues showed that the final partitionings have the same "variation of information" (in the sense of Meilă 2007) as languages from the World Color Survey with the same number of color terms. Visual inspection of the partitionings that resulted after thirteen generations suggests that these partitionings also fulfill several of the design criteria that we have presented. For instance, while Xu and colleagues did not consider the Convexity constraint, the color partitionings of the five learning chains they present in their Figure 2 (Xu et al. 2013, 4) show clear signs of convexity already after four generations of learning. In a related experiment involving ten generations of learners, Carstensen et al. (2015) obtained similar results concerning spatial relations based on the Topological Re-

[15] For a similar result concerning artificial agents, see Steels and Belpaeme (2005).

lations Picture Series (Bowermann & Pederson 1992). These authors also showed that the partitionings become increasingly informative over the generations, where informativeness was measured as in Regier, Kay, and Khetarpal (2007).[16] And re-analyzing Xu et al.'s results, Carstensen et al. found the same increasing informativeness in those results, indicating that they satisfy WELL-FORMEDNESS.

The simulations and experiments discussed here are evidence that the design criteria and constraints proposed in previous sections are instrumental in facilitating efficient communication of concepts. Hence, our proposal receives support from both individualistic and social considerations.

7 Conclusion

We started with an open theoretical question for the conceptual spaces framework, namely, the question of which regions in conceptual spaces represent or could represent *natural* concepts. A preliminary answer offered in Gärdenfors' work—natural concepts are represented by *convex* regions—was seen not to suffice. We proposed a different answer in terms of optimal design: natural concepts are represented by the cells of an optimally designed similarity space, where we defined the notion of optimal design by reference to a number of general principles. These principles were motivated from an engineering perspective that took into account our cognitive limitations as well as the world we inhabit, as individuals and as a community of interacting agents.

Although we have not explicitly discussed the use of conceptual systems in artificial systems such as robots, it is clear that the design perspective we are proposing here has implications for how concept learning in artificial systems can be implemented, in particular systems that are used in communication with humans.[17] We described various computational procedures for categorization, both in static and in dynamic contexts, that all were seen to implement several of our design principles. That these procedures then yielded categorizations remarkably similar to ones lexicalized by various spoken languages was taken as evidence that those design principles are operative also in human categorization.

Much of the more detailed support for our proposal came from a rather limited domain, to wit, that of color concepts. To an important extent, this is due to the fact that few other domains have been explored as thoroughly by cognitive psychologists. As Clark (1993, vii) remarks, color research is "*the* success story of scientific psychology thus far."[18] Indeed, we cannot think of a second field of research that offers data on categorization as rich as those available through the World Color Survey, mentioned in Section 6. That being said, a greater variety of evidence is certainly needed to reach a more definite verdict on the idea that design plays a key role in shaping our concepts *generally*. At the outset of Section 2, we listed a number of conceptual spaces besides color space that have drawn the attention from researchers, and the obvious way to obtain further evidence for our proposal is to gather more categorization data pertinent to those spaces and see whether and to what extent they confirm that design principles are operative in forming concepts. The work on

[16] See Kemp and Regier (2012), Xu and Regier (2014), and Xu, Regier, and Malt (2016) for similar findings regarding kinship categories, numerical systems, and container categories, respectively.

[17] A design perspective is to be found in much recent work on biologically-inspired cognitive architectures; see Lieto et al. (2018) and references given there.

[18] See in the same vein various chapters in Elliot, Fairchild, and Franklin (2015).

color discussed in the present paper suggests a variety of ways in which such future work could be fruitfully conducted.

Another route to take, apart from the psychological one, is to employ methods from machine learning, including neural networks, to generate similarity spaces that can form the foundation for the design of conceptual systems. Methods for handling "big data" such as deep learning (LeCun, Bengio, & Hinton 2015; Goodfellow, Bengio, & Courville 2016) promise to summarize the data from semantic domains that can be represented in similarity spaces. In this way, the cumbersome collecting of human similarity judgments may be circumvented and replaced by more automatized techniques.[19]

References

Abramov, I. (1997) Physiological mechanisms of color vision. In C. L. Hardin & L. Maffi (eds.) *Color categories in thought and language*, pp. 89–117. Cambridge: Cambridge University Press.

Abramov, I. & Gordon, J. (1994) Color appearance: On seeing red—or yellow, or green, or blue. *Annual Review of Psychology* 45:451–85.

Aisbett, J. & Gibbon, G. (2001) A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence* 133:189–232.

Alon, U. (2003) Biological networks: The tinkerer as an engineer. *Science* 301:1866–67.

Aydede, M. & Guzeldere, G. (2005) Concepts, introspection, and phenomenal consciousness: An information-theoretic approach. *Noûs* 39:197–255.

Bååth, R., Lagerstedt, E., & Gärdenfors, P. (2014) A prototype-based resonance model of rhythm categorization. *i-Perception* 5:548–58.

Berlin, B. & Kay, P. (1969/1999) *Basic color terms.* Stanford CA: CSLI Publications.

Borg, I. & Groenen, P. (2010) *Modern multidimensional scaling* (2nd ed.). New York: Springer.

Bosten, J. M., Robinson, J. D., Jordan, G., & Mollon, J. D. (2005) Multidimensional scaling reveals a color dimension unique to "color-deficient" observers. *Current Biology* 15:R950–52.

Bowerman, M. & Pederson, E. (1992) Topological relations picture series. In S. C. Levinson (ed.) *Space stimuli kit 1.2* (November 1992, 51). Nijmegen: Max Planck Institute for Psycholinguistics.

Brössel, P. (2017) Rational relations between perception and belief: The case of color. *Review of Philosophy and Psychology* 8:721–41.

Carey, S. (2009) *The origin of concepts.* Oxford: Oxford University Press.

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015) Language evolution in the lab tends toward informative communication. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin TX: Cognitive Science Society.

[19] We are greatly indebted to Lieven Decock, Yasmina Jraissati, Christopher von Bülow, and two anonymous referees for valuable comments on previous versions of this paper. We are also grateful to audiences in Bochum, Düsseldorf, Paris, Regensburg, and Stockholm.

Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013) Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE* 8:e73289. `https://doi.org/10.1371/journal.pone.0073289`.

Chella, A., Frixione, M., & Gaglio, S. (1997) A cognitive architecture for artificial vision. *Artificial Intelligence* 89:73–111.

Chella, A., Frixione, M., & Gaglio, S. (2000) Understanding dynamic scenes. *Artificial Intelligence* 123:89–132.

Chella, A., Frixione, M., & Gaglio, S. (2003) Anchoring symbols to conceptual spaces: The case of dynamic scenarios. *Robotics and Autonomous Systems* 43:175–88.

Churchland, P. M. (1986) Some reductive strategies in cognitive neurobiology. *Mind* 95:279–309.

Churchland, P. M. (2012) *Plato's camera*. Cambridge MA: MIT Press.

Clark, A. (1993) *Sensory qualities*. Oxford: Oxford University Press.

Cook, R. S., Kay, P., & Regier, T. (2005) The World Color Survey database: History and use. In H. Cohen & C. Lefebvre (eds.) *Handbook of categorization in cognitive science*, pp. 223–42. Amsterdam: Elsevier.

Decock, L. & Douven, I. (2014) What is graded membership? *Noûs* 48:653–82.

Decock, L., Douven, I., Kelp, C., & Wenmackers S. (2014) Knowledge and approximate knowledge. *Erkenntnis* 79:1129–50.

Derrac, J. & Schockaert, S. (2015) Inducing semantic relations from conceptual spaces. *Artificial Intelligence* 228:66–94.

De Valois, R., Abramov, I., & Jacobs, G. H. (1966) Analysis of response patterns of LGN cells. *Journal of the Optical Society of America* 56:966–77.

De Valois, R. & De Valois, K. (1993) A multi-stage color model. *Vision Research* 36:833–36.

De Valois, R., De Valois, K., & Mahon, L. (2000) Contribution of S opponent cells to color appearance. *Proceedings of the National Academy of Sciences USA* 97:512–17.

Douven, I. (2016) Vagueness, graded membership, and conceptual spaces. *Cognition* 151:80–95.

Douven, I. (2017) Clustering colors. *Cognitive Systems Research* 45:70–81.

Douven, I. (2018a) The rationality of vagueness. In R. Dietz (ed.) *Vagueness and Rationality*, New York: Springer, forthcoming.

Douven, I. (2018b) New foundations for fuzzy set theory. In A. Aberdein & M. Inglis (eds.) *Advances in experimental philosophy*, London: Bloomsbury, forthcoming.

Douven, I. & Decock, L. (2017) What verities may be. *Mind* 126:386–428.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013) Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic* 42:137–60.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017) Measuring graded membership: The case of color. *Cognitive Science* 41:686–722.

Du, K.-L. (2010) Clustering: A neural network approach. *Neural Networks* 23:89–107.

Elliot, A. J., Fairchild, M. D., & Franklin, A. (2015) *Handbook of color psychology*. Cambridge: Cambridge University Press.

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

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

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

Gärdenfors, P. (2007) Representing actions and functional properties in conceptual spaces. In T. Ziemke, J. Zlatev, & R. M. Frank (eds.) *Body, language and mind* (Vol. 1), pp. 167–95. Berlin: De Gruyter.

Gärdenfors, P. (2014) *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge MA: MIT Press.

Gärdenfors, P. & Warglien, M. (2012) Using concept spaces to model actions and events. *Journal of Semantics* 29:487–519.

Gärdenfors, P. & Zenker, F. (2011) Using conceptual spaces to model the dynamics of empirical theories. In E. J. Olsson & S. Enqvist (eds.) *Belief revision meets philosophy of science*, pp. 137–53. New York: Springer.

Gärdenfors, P. & Zenker, F. (2013) Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese* 190:1039–58.

Gini, C. (1921) Measurement of inequality of incomes. *Economic Journal* 31:124–26.

Goodfellow, I., Bengio, Y., & Courville, A. (2016) *Deep learning*. Cambridge MA: MIT press.

Goodman, N. (1954) *Fact, fiction, and forecast*. London: Athlone Press.

Green, S., Levy, A., & Bechtel, W. (2015) Design sans adaptation. *European Journal for Philosophy of Science* 5:15–29.

Indow, T. (1988) Multidimensional studies of Munsell color solid. *Psychological Review* 95:456–70.

Jäger, G. (2007) The evolution of convex categories. *Linguistics and Philosophy* 30:551–64.

Jäger, G. (2010) Natural color categories are convex sets. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (eds.) *Logic, language and meaning*, pp. 11–20. Berlin: Springer.

Jäger, G. & van Rooij, R. (2007) Language structure: Psychological and social constraints. *Synthese* 159:99–130.

Jakobson, R. (1968) *Child language, aphasia, and phonological universals*. The Hague: Mouton.

Jameson, K. A. & D'Andrade, R. G. (1997) It's not really red, green, yellow, blue. In C. L. Hardin & L. Maffi (eds.) *Color categories in thought and language*, pp. 295–319. Cambridge: Cambridge University Press.

Johannesson, M. (2000) Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology* 53:121–39.

Johnson, K. (2008) *Quantitative methods in linguistics*. Oxford: Blackwell.

Jraissati, Y. & Douven, I. (2017) Does optimal partitioning of color space account for universal color categorization? *PLoS ONE* 12: e0178083. `https://doi.org/10.1371/journal.pone.0178083`.

Jraissati, Y. & Douven, I. (2018) Delving deeper into color space. *i-Perception*, forthcoming.

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

Kaufman, L. & Rousseeuw, P. J. (1990) *Finding groups in data*. Hoboken NJ: Wiley.

Kay, P., Berlin, B., & Merrifield, W. (1991) Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology* 1:12–25.

Kay, P. & McDaniel, C. K. (1978) The linguistic significance of the meaning of basic color terms. *Language* 54:610–46.

Kemp, C. & Regier, T. (2012) Kinship categories across languages reflect general communicative principles. *Science* 336:1049–54.

Kriegeskorte, N. & Kievit, R. A. (2013) Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* 17:401–12.

Kriegeskorte, N. & Mur, M. (2012) Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology* 3:245. `https://doi:10.3389/fpsyg.2012.00245`.

Krumhansl, C. L. (1978) Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review* 85:445–63.

Kullback, S. & Leibler, R. A. (1951) On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.

Lakoff, G. (1973) Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* 2:458–508.

Lawry, J. & Eyre, H. (2014) Language games with vague categories and negations. *Adaptive Behavior* 22:289–303.

Lawry, J. & Tang, Y. (2009) Uncertainty modelling for vague concepts: A prototype theory approach. *Artificial Intelligence* 173:1539–58.

LeCun, Y., Bengio, Y., & Hinton, G. (2015) Deep learning. *Nature* 521:436.

Lewis, D. K. (1983) New work for a theory of universals. *Australasian Journal of Philosophy* 61:343–77.

Lewis, M. & Lawry, J. (2014) A label semantics approach to linguistic hedges. *International Journal of Approximate Reasoning* 55:1147–63.

Lewis, M. & Lawry, J. (2016) Hierarchical conceptual spaces for concept combination. *Artificial Intelligence* 237:204–27.

Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018) The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research* 48:1–3.

Lieto, A., Chella, A., & Frixione, M. (2017) Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures* 19:1–9.

Lieto, A., Minieri, A., Piana, A., & Radicioni, D. P. (2015) A knowledge-based system for prototypical reasoning. *Connection Science* 27:137–52.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48:839–62.

Machery, E. (2009) *Doing without concepts.* Oxford: Oxford University Press.

Malacara, D. (2002) *Color vision and colorimetry: Theory and applications.* Bellingham WA: SPIE Press.

Malt, B. C. (1994) Water is not H_2O. *Cognitive Psychology* 27:41–70.

Malt, B. C. (1995) Category coherence in cross cultural perspective. *Cognitive Psychology* 29:85–148.

Malt, B. C. & Sloman, S. A. (2007) Category essence or essentially pragmatic? Creator's intention in naming and what's really what. *Cognition* 105:615–48.

Malt, B. C., Sloman, S. A., & Gennari, S. (2003) Universality and language specificity in object naming. *Journal of Memory and Language* 49:20–42.

Margolis, E. & Laurence, S. (2007) The ontology of concepts: Abstract objects or mental representations? *Noûs* 41:561–593.

Marr, D. C. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Medin, D. L. (1989) Concepts and conceptual structure. *American Psychologist* 44:1469–81.

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000) Are there kind of concepts? *Annual Review of Psychology* 51:121–47.

Meilă, M. (2007) Comparing clusterings: An information based distance. *Journal of Multivariate Analysis* 98:873–95.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013) Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology* 4:128. `https://doi=10.3389/fpsyg.2013.00128`.

Murphy, G. L. (2002) *The big book of concepts*. Cambridge MA: MIT Press.

Nosofsky, R. M. (1987) Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:87–108.

Nowak, M. A. (2006) *Evolutionary dynamics: Exploring the equations of life*. Cambridge MA: Harvard University Press.

Nussbaum, C. (2015) Musical perception. In M. Matthen (ed.) *The Oxford handbook of philosophy of perception*, pp. 495–514. Oxford: Oxford University Press.

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000) *Spatial tessellations* (2nd ed.). New York: Wiley.

Osherson, D. N. & Smith, E. E. (1981) On the adequacy of prototype theory as a theory of concepts. *Cognition* 9:35–58.

Osherson, D. N. & Smith, E. E. (1982) Gradedness and conceptual combination. *Cognition* 12:299–318.

Osherson, D. N. & Smith, E. E. (1997) On typicality and vagueness. *Cognition* 64:189–206.

Peterson, M. (2017) *The ethics of technology: A geometric analysis of five moral principles*. Oxford: Oxford University Press.

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

Poyatos, J. F. (2012) On the search for design principles in biological systems. *Advances in Experimental Medicine and Biology* 751:183–93.

Regier, T., Kay, P., & Khetarpal, N. (2007) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences USA* 104:1436–41.

Regier, T., Kay, P., & Khetarpal, N. (2009) Color naming and the shape of color space. *Language* 85:884–92.

Regier, T., Kemp, C., & Kay, P. (2015) Word meanings across languages support efficient communication. In B. MacWinnhey & W. O'Grady (eds.) *The handbook of language emergence*, pp. 237–63. Hoboken NJ: Wiley.

Rey, G. (1983) Concepts and stereotypes. *Cognition* 15:237–62.

Rey, G. (1985) Concepts and conceptions: A reply to Smith, Medin and Rips. *Cognition* 19:297–303.

Rosch, E. (1973) Natural categories. *Cognitive Psychology* 4:328–50.

Roy, D. (2005a) Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence* 167:170–205.

Roy, D. (2005b) Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences* 9:389–96.

Savageau, M. A. (2001) Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos* 11:142–59.

Schockaert, S. & Prade, H. (2013) Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. *Artificial Intelligence* 202:86–131.

Shepard, R. N. (1964) Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology* 1:54–87.

Shepard, R. N. (1982) Geometrical approximations to the structure of musical pitch. *Psychological Review* 89:305–33.

Simon, H. A. (1972) Theories of bounded rationality. In C. B. McGuire & R. Radner (eds.) *Decision and organization*, pp. 161–76. Amsterdam: North-Holland.

Sivik, L. & Taft, C. (1994) Color naming: A mapping in the IMCS of common color terms. *Scandinavian Journal of Psychology* 35:144–64.

Steels, L. (ed.) (2012) *Experiments in cultural language evolution*. Amsterdam: John Benjamins.

Steels, L. & Belpaeme, T. (2005) Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences* 28:469–88.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016) Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology* 69:1996–2019.

Warglien, M. & Gärdenfors, P. (2013) Semantics, conceptual spaces, and the meeting of minds. *Synthese* 190:2165–93.

Xu, J., Dowman, M., & Griffiths, T. (2013) Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B* 280:20123073.

Xu, Y. & Regier, T. (2014) Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (eds.) *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin TX: Cognitive Science Society.

Xu, Y., Regier, T., & Malt, B. C. (2016) Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science* 40:2081–94.

Zwarts, J. & Gärdenfors, P. (2016) Locative and directional prepositions in conceptual spaces: The role of polar convexity. *Journal of Logic, Language and Information* 25:109–38.