Similarity Space Theories and the Problem of Concept Acquisition

This page intentionally left blank

Similarity Space Theories and the Problem of Concept Acquisition

PhD Thesis in Philosophy
2019

José Vicente Hernández Conde
University of the Basque Country

Supervisor
Professor Agustín Vicente Benito

This page intentionally left blank

For Julia and Julia

This page intentionally left blank

Declaration

I hereby declare that this dissertation is my own original work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due acknowledgement is made in the text. I also declare that no part of this thesis has been previously submitted, either in the same or different form, for the award of any other degree at this or any other university.

José Vicente Hernández Conde

This page intentionally left blank

Summary

Abstract

One of the main problems of concept empiricism is to explain the acquisition of the most basic constituents of concepts, without resorting to preexisting innate elements. The aim of this thesis is to show that the best nativist arguments against the acquisition of (primitive) concepts rest on the assumption that the constituents of concepts should be available beforehand, as an input of the acquisition process. However, I will claim that there is no obligation to accept such a (precedence) assumption. In fact, I will describe a model where the constituents of a concept result from the same learning process by virtue of which that concept is acquired. My proposal is based on a similarity space theory of concepts articulated by means of prototypes. I also prove that: (A) in this type of approach, two distinct notions of concept should be distinguished –which may be identified with two different facets in their life cycle (i.e., storage and instantiation)–; and that (B) a proposal like this brings together virtues both from the invariantist and from the contextualist views. I argue as well that, if concepts are context-dependent, as claimed by contextualism, then instantiated concepts lack minimal persistence and, consequently, cannot be a representation of their associated categories.

Resumen

Uno de los principales problemas a los que se enfrenta el empirismo es el de explicar cómo se adquieren los elementos más básicos de los conceptos, sin recurrir para ello a elementos innatos preexistentes. El propósito de esta tesis es mostrar que los mejores argumentos nativistas en contra de la posibilidad de aprender conceptos (primitivos) dependen de la asunción de que los constituyentes de los conceptos deben estar disponibles de antemano, como entrada de los procesos de adquisición. No obstante, mostraré que nada obliga a aceptar esa asunción (de precedencia). De hecho, presentaré un modelo en donde los elementos constitutivos de un concepto resultan del mismo proceso de aprendizaje en virtud del cual ese concepto se adquiere. Mi propuesta está basada en una teoría de espacios de similaridad articulada mediante prototipos. Además pruebo: (A) que dos nociones distintas de concepto deben distinguirse en este tipo de aproximación, a saber, conceptos como almacenamiento y conceptos como instanciación; y (B) que una propuesta como ésta reúne virtudes tanto del ámbito invariantista como del contextualista. Argumento también que, si los conceptos son dependientes del contexto –según sostiene el contextualismo–, entonces los conceptos instanciados carecen de persistencia mínima y, por ello, no pueden ser una representación de sus categorías asociadas.

Publications

Articles in Journals

- Forthcoming. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology* (first-authored by Florian Cova, co-authored by the XPhi Replicability Project).
- Forthcoming. Articulating context dependence. (Accepted for publication as part of a Special Issue; publisher still unknown.)
- 2017. A case against convexity in conceptual spaces. *Synthese* 194: 4011-4037.
- 2017. Life cycle of a concept in the *ad hoc* cognition framework. *Theoria. An International Journal for Theory, History and Foundations of Science* 32: 271-292.

Conference Proceedings

- 2018. Cuando los conceptos son no-persistentes y no-representacionales. In C. Saborido, S. Oms and J. González de Prado, eds., *Proceedings of the IX Conference of the Spanish Society for Logic, Methodology and Philosophy of Science*. Madrid: UNED, 226-229.
- 2015. Characterization of antonyms in conceptual spaces. In J. Díez, M. García-Carpintero, J. Martínez and S. Oms, eds., *Proceedings of the VIII Conference of the Spanish Society for Logic, Methodology and Philosophy of Science*. Barcelona: University of Barcelona, 128-133.

This page intentionally left blank

Acknowledgments

First of all, my greatest debt in writing this thesis is owed to my supervisor, Agustín Vicente, whose encouragement, intellectual guidance and unwavering support through difficult circumstances has been steady throughout all these years. His mentorship was crucial everywhere in the lengthy process of defining the research project and carrying it to completion. Besides, his detailed and patient readings of my work, and his insightful criticisms of many of my ideas, have greatly improved this thesis. He has taught me how to do proper research, and his dedication has allowed focusing my efforts with countless suggestions that have led to deepen and refine varied aspects of my thought. Agustín has been the supervisor that most graduate students can only dream of, and I feel truly privileged to have worked under his direction.

At this point, I wish to declare my indebtedness to the Department of Philosophy at the University of Valladolid, because many ideas in this dissertation are, one way or another, a result of the years spent there. Special thanks are due to my teachers Cristina Corredor, Juan Barba, María Caamaño and Maximiliano Fartos. They were always kind and patient enough to talk with me about any issue, and their useful suggestions, and warm and inspiring stances have extended up until now. I am indebted to Cristina Corredor, whose advice led me to decide to carry out my PhD under the supervision of Agustín Vicente. Thanks to Juan Barba as well, for sharing so many fascinating discussions, whose enlightening view on logic has steadily stimulated me. My gratitude goes also to María Caamaño and Maximiliano Fartos, whose voices have been a source of guidance and perspective all through these years.

I would also wish to express my gratitude to the University of the Basque Country, which awarded me with a FPI scholarship in the year 2013, and provided the financial support for my stays as visiting scholar at the Universities of Salzburg and Pittsburgh. I am indebted to the Spanish Ministry of Economy and Competitiveness as well, which – through the research projects FFI2011-30074-C02-02 and FFI2014-52196-P– provided funding for attending to numerous conferences and workshops around Europe.

In the spring and summer of 2015 I spent three months as a visiting researcher at the University of Salzburg. Big thanks to Christopher Gauker for accepting my visit and allowing me to participate in all the activities of the Philosophy Department. I am profoundly grateful to Chris for the helpful discussions on my views, and useful advice on how to express them. His insightful feedback on my ideas made a substantial contribution to my PhD project. During this stay, my work also profited from conversations with Johannes Brandl and Julien Murzi. I must also mention Alexander Hieke, whose warm welcome and hospitality contributed hugely to the stimulating and enjoyable experience that was the time expended in Salzburg.

My second stay abroad, in the fall of 2016, was in the Department of History and Philosophy of Science at the University of Pittsburgh. I owe a great thanks to Edouard Machery, my host advisor, for his guidance and making my visit possible. I am in huge debt to him for reading and giving detailed comments on several drafts of my papers, and for the fruitful discussions which brought up some important issues of my work. His endless energy and excellence were a continuous inspiration to me. During those days in Pittsburg, I also benefited from the superlative intellectual environment of the Center for the Philosophy of Science, and from conversations with Anil Gupta, Kareem Khalifa, and Anjan Chakravartty. In that period I was fortunate to begin the collaboration with the XPhi Replicability Project as well, which has been a particularly valuable experience.

During this time, I have had the opportunity to present the ideas put forward in this doctoral thesis in many venues, where I have benefited from the comments, questions and criticisms of the audience members in Madrid, Barcelona, Valencia, Nijmegen, Cambridge, Warsaw, Södertörn, Düsseldorf, Munich and Osnabrück. The talks and discussions with numerous philosophers in those meetings contributed greatly to improve the ideas in this dissertation. I wish to thank them all as well.

Last but not least, I would like to mention the many people I have met at the University of the Basque Country throughout this time. I am thankful to the members of the different reading groups organized around Agustín Vicente, for the insightful talks and discussions. Many thanks go to Andoni Ibarra, for his unceasing encouragements and help with numerous bureaucratic issues. Thanks also to Dora Martínez, without whose assistance in paperwork things would have been much more tedious.

Finally, I am very grateful to my family, particularly to my mother and brother, Eusebia and Juan, and to my uncle Luciano, for their company and continuous support. *Gracias*. And, most importantly, I dedicate this PhD thesis to my wife, Julia, a loving companion of extraordinary patience and kindness. She was always comprehensive and supportive through the breakdowns and stressful times in the realization of this work. This thesis is also dedicated to my daughter, Julia, who arrived at the initial period of the writing.

Contents

Figures

Tables

This page intentionally left blank

Similarity Space Theories and the Problem of Concept Acquisition

This page intentionally left blank

Introduction

My purpose is to better understand human nature. My method is to attempt to characterize the mental resources that make possible the articulation of humans' knowledge and experience of the world. –
Ray S. Jackendoff (1989, pp. 68-69)

Achilles had overtaken the Tortoise, and had seated himself comfortably on its back[1].

"I have got to the end of my racecourse," said Achilles.

"That is a remarkable claim," said the Tortoise. "Even though I do not know what a *racecourse* is."

"What a surprising question," replied Achilles. "I cannot believe you do not know it. In fact, that is an extremely easy issue which I learned a long time ago: a *racecourse* is the *cour*se of a *race*."

"So you learned in the past what a *racecourse* is," the Tortoise commented.

"Quite so," Achilles agreed.

"And, if I have correctly understood, the concept of *racecourse* is constituted by the concepts of *race* and *course*, so once you know what a *course* is, and what a *race* is, you know what a *racecourse* is."

"That is the beauty of the compositionality principle," Achilles assented.

"But, if you had not known what a *race* is –or what a *course* is– then you could not have learned what a *racecourse* is," the Tortoise wondered.

"Undoubtedly! But, at that time, I had already learned what a *race* and a *course* are," retorted Achilles. "And I am sure you know what they are too."

"Well, now let's take a look at the argument," the Tortoise interrupted. "Let's refer *racecourse*, *race* and *course* by means of letters, *A*, *B*, and *C*, respectively. So, when you learned *A* you had already learned *B* and *C*."

"Yes, of course," answered Achilles.

[1] This is the beginning of Lewis Carroll's paper "What the Tortoise said to Achilles" (Carroll 1895), and the rest of the dialogue is inspired by that text.

"Then, I must ask you how you learned the concepts *B* and *C*," said the Tortoise. "Let's begin with the concept *C*."

"Through the same process," Achilles replied. "I learned it at some previous time."

"Now I understand," the Tortoise murmured. "You learned the concept *C* on the basis of other concepts, let's call them *D* and *E*, that you knew at that time."

"I did!", exclaimed Achilles.

"But, in such a case, I wonder how you learned the concepts *D* and *E*," the Tortoise said musingly. "Did you learn *D* and *E* from other concepts that you had learned at an earlier time?"

"I see," Achilles muttered; *and there was a touch of sadness in his tone*.

This example illustrates the main issue I investigate in this doctoral thesis. Empiricism is one of the two major views on the origin of concepts. According to it, concepts are learned from sense experience, and few if any of them are innate. More particularly, if the principle of compositionality is accepted, as usual in contemporary philosophy, we arrive at Achilles' position in the dialogue above.

Unfortunately, concept empiricism faces a significant problem when trying to explain how the most basic constituents of concepts could have been acquired. That was the Tortoise's point in the dialogue (i.e., there is a circularity threat whenever concepts and their constitutive elements are thought to be acquired by means of the same kind of cognitive process), and the roots of Fodor's nativist critiques against the thesis that primitive concepts can be learned without resorting to a preexisting innate repertoire of concepts. But, things might get even worse for the empiricist, since if Fodor is right when claiming that all available concepts are the closure of the primitive ones under a set of innate combinatorial mechanisms, then the expressive power of all our conceptual system could be innately determined.

However, the empiricist-nativist debate is not the only great discussion in current philosophy regarding the nature of concepts. In fact, other significant debate is the one about the invariant or context-dependent character of concepts. This is the other major topic of this doctoral thesis. According to invariantism, concepts are stable bodies of knowledge about categories, which remain invariant across individuals and time. In contrast, the contextualist view holds that concepts are context-dependent construals produced on the fly for each particular occasion. Each one of these views accounts for several important kinds of phenomena. On the one hand, invariantism easily explains the accumulation of knowledge –by subjects– about categories, and our ability to communicate with other individuals. On the other hand, contextualism accounts for our adaptive behavior to heterogeneous and changing environments.

With regard to this second issue, my proposal will be that a similarity-based space theory of concepts articulated by means of prototypes can bring together virtues both from the invariantist and from the contextualist views. In particular, I will claim that an approach like mine allows to distinguish two different notions of concept, namely stored concept and instantiated concept, which may be associated with distinct facets in the life cycle of a concept: (a) *stored concepts* would contain the information needed to be persistently kept by the mind about a concept for its subsequent instantiation; (b) *instantiated concepts* would be the result of those cognitive process where the concept is applied (i.e., categorizations, inferences, etc.) By virtue of this, stored concepts will be able to explain many typically invariantist phenomena; while instantiated concepts, which are produced in a context-dependent way, will account for those aspects commonly explained by contextualism.

In respect to the circularity threat on concept empiricism, I will show that the uppermost nativist arguments against the acquisition of primitive concepts rely on the assumption that the constituents of a concept *C* must be available as an input of the learning process by virtue of which the concept *C* is acquired. Notwithstanding, I will claim that there is no obligation to accept such a hypothesis. My proposal will be that a model where the constitutive elements of a concept result from the same acquisition process which leads to the learning of that concept may explain the formation of concepts in a non-circular way. The proposed model will consist in a three-step iterative learning system, constituted by two general-purpose abilities (i.e., dimensional reduction and pattern identification), and one final stage of evaluation and readjustment of the model. First, the dimensional reduction will produce new reduced factors –by ruling out as much redundant information as possible–, which might be identified with the most basic elements of our conceptual system. Secondly, the pattern recognition stage would search for regularities within the reduced data; such regularities could be identified with the concepts of our mental system. Finally, an iterative process is necessary since nothing guarantees that the obtained factors and patterns are the most predictive ones, and that is also the reason for the third stage of the process, where the model is evaluated and readjusted.

Organization of the thesis

Chapters 1, 2, and 3 are introductory chapters. Chapter 1 describes the different theories on the nature, origin, internal structure, and contextual dependence of concepts. There I examine the distinct approaches, and the relations existing between the different views. In this chapter I also make explicit my presumptions on these issues, and the reasons for them. In particular, I will opt for an empiricist-contextualist perspective, which will not be committed to any specific theory on the nature of concepts.

Chapter 2 is focused on the major possible views about the structure of concepts since, even for those who think that concepts have internal structure, there is controversy concerning which type of conceptual structure is more appropriate to characterize them. Those main views are theories based on definitions (classical theory), on similarities (prototype and exemplar theories), and on explanations (theory theory); as well as those other approaches (i.e. atomism, hybridism, pluralism, and eliminativism) which emerged in response to the inability of the previous theories to provide a complete and successful explanation of the main empirical phenomena.

Chapter 3 deals with the similarity-based approaches to the structure of concepts. Because there is no single way of characterizing the idea of similarity, but rather a wide range of possible similarity models and measures, the aim of this chapter is to examine the four main contemporary models of similarity (i.e., geometric models, featural models, alignment-based models and transformational models), in order to clarify to what extent they can explain the observed phenomena.

Chapter 4 is devoted to the notion of conceptual space, understood as a framework for the representation of concepts and knowledge. In the first part I describe what a similarity space theory of concepts is and, in particular, I focus on Gärdenfors' conceptual spaces. In the second part, I discuss the role played by convexity in this latter approach, and there I argue that Gärdenfors' convexity constraint –according to which conceptual regions should be convex– is unnecessary from a theoretical perspective, and problematic with regard to some particular applications of his theory. My conclusion will be that, if the convexity criterion is abandoned, then Gärdenfors' theory can be reduced to a contextualist geometric articulation of the prototype theory.

Chapter 5 investigates how a prototype theory articulated by means of a similarity-based conceptual space could bring together virtues both from invariantism and from contextualism. In particular, I will show that Casasanto and Lupyan's *ad hoc* cognition framework, according to which all concepts are produced *ad hoc* when they are instantiated –and so, there are no context-independent concepts–, may be characterized by means of a similarity-based theory of concepts. On this basis I will show that two different notions of concept should be distinguished, which may be identified with two distinct facets in their life cycle (i.e., storage and instantiation). Lastly, I will argue for the thesis that, if concepts are presumed to be context-dependent –as claimed by contextualists–, then instantiated concepts lack minimal persistence and, in consequence, cannot be a representation of their associated categories.

Chapter 6 examines how it can be explained the acquisition of the most basic constituents of concepts from an empiricist point of view, without resorting to an innate repertoire of elements. In this chapter I prove that the best nativist arguments against concept empiricism depend on what I call the *precedence assumption*, that is, the hypothesis that the constituents of concepts must be available beforehand the beginning of their respective learning processes. Then, I sketch out a model which is able to produce both concepts and their constitutive properties as result of the same execution of the acquisition process. This refutes nativist arguments against empiricism because, once the precedence assumption is surpassed, circularity is no longer a threat for the learning of primitive concepts.

Capítulo 1: Conceptos

Nihil est in intellectu quod prius non fuerit in sensu. –
Tomás de Aquino (1256-1259, II iii 19)[1]

... excipe: *nisi ipse intellectus.* –
Gottfried W. Leibniz (1765, II i 2)

Categorizaciones, inferencias, generalizaciones, predicciones, aprendizaje, memoria, toma de decisiones, comunicación, resolución de problemas son –todos ellos– fenómenos cognitivos en donde la noción de concepto desempeña un papel fundamental. De hecho, en prácticamente cualquier actividad humana que examinemos lo que encontramos son sujetos categorizando, generalizando, reconociendo semejanzas, realizando inferencias o tomando decisiones en base a tales semejanzas y categorizaciones, etc., y todas estas actividades las concebimos como soportadas por aquello a lo que llamamos *conceptos*.

Conforme indica Murphy (2002), puesto que raras veces nos encontramos ante una misma entidad –esto es, el mismo perro o el mismo árbol–, dependemos de nuestros conceptos del mundo para comprender lo que sucede a nuestro alrededor. A saber, categorizamos bajo categorías conocidas cosas vistas por primera vez, y atribuimos propiedades observadas en ejemplos conocidos de una cierta categoría a aquellas nuevas entidades a las que clasificamos en ese grupo. Dicho de otro modo, necesitamos los conceptos para conectar experiencias pasadas con nuestra experiencia actual, para así poder determinar lo que algo es, y qué propiedades tiene. Aún más, no solo somos capaces de formarnos conceptos acerca del mundo que percibimos de modo inmediato, sino también de entidades espacial o temporalmente lejanas (como el planeta Venus o Julio César), abstractas (como justicia o los números), e incluso sin existencia (como las hadas o el éter).

No obstante, a pesar del importante papel atribuido a la noción de concepto, no existe un consenso general con respecto a su naturaleza, adquisición y estructura. Así, aunque en la filosofía de la mente los conceptos suelen ser presentados como los constituyentes

[1] Adaptación de una frase de Aristóteles.

últimos de los pensamientos[2] (Rey 1994; Solomon, Medin y Lynch 1999; Margolis y Laurence 2003; 2011a), hay también ocasiones en que son concebidos como principios –o dispositivos– de categorización, esto es, como algo que permite determinar si una cierta entidad pertenece o no a la categoría considerada (Price 1953; Geach 1957; Prinz 2002), e incluso como meros cuerpos de conocimiento sobre los miembros de una determinada categoría (Barsalou 1993; Machery 2009). Y no menos controvertida es la cuestión relativa a cuál puede ser la estructura interna de las representaciones mentales con que –en la filosofía de la mente– se identifican los conceptos (Medin 1989; Komatsu 1992), y que, principalmente, los conciben como definiciones, prototipos, ejemplares o teorías.

Finalmente, casi todas las propuestas toman alguna posición en cuanto a la naturaleza y estructura interna de los conceptos, así como con respecto a ciertos debates principales –como, por ejemplo, la discusión empirista-nativista sobre el origen de los conceptos, o el modo en que éstos dependen del contexto–. El propósito de este capítulo es presentar las posturas más generales que pueden adoptarse con respecto a la estructura interna[3], naturaleza, origen y grado de dependencia contextual de los conceptos.

1.1. Estructura interna de los conceptos

Una de las distinciones más básicas que suelen establecerse en el ámbito de los conceptos es la que diferencia entre conceptos primitivos y complejos. Los *conceptos primitivos*, también llamados conceptos atómicos, serían aquellos que no tienen estructura interna, esto es, que no están constituidos por otros conceptos. Por contraposición, los *conceptos complejos* serían aquellos que no son primitivos[4].

Así, por ejemplo, el concepto léxico[5] *complejo* SOLTERO estaría constituido por otros conceptos más simples –tales como NO-CASADO, VARÓN y ADULTO–, los cuales estarían a su vez constituidos por otros conceptos aún más simples, hasta llegar a un punto en que esos elementos constitutivos fueran conceptos primitivos. Obviamente, esto no sería

[2] La visión de que los conceptos son los elementos constitutivos (*building blocks*) de los pensamientos se remonta, al menos, hasta Frege (1914, p. 225), aunque Davidson (1977, p. 252) considera que las teorías de los *building blocks* –en el caso de la semántica– están ya presentes en el empirismo británico (Berkeley, Hume, Mill).

[3] El capítulo 2 estará específicamente dedicado a la cuestión de cuáles son las principales teorías sobre la estructura de los conceptos –a saber, teoría clásica, prototipos, ejemplares, teoría-teoría, atomismo, hibridismo, pluralismo y eliminativismo–, y a la discusión de sus más importantes puntos fuertes y débiles.

[4] Ésta es justamente la diferencia entre las ideas –o conceptos– simples y complejas del empirismo británico, según el cual las ideas complejas no proceden de la experiencia, sino que son construidas a partir de ideas simples que sí tendrían su origen en la experiencia (Locke 1690: II ii 1-2; Hume 1741 cap. II).

[5] Los *conceptos léxicos* son conceptos asociados con ítems léxicos de los lenguajes naturales (Laurence y Margolis 1999, p. 4) –como lo son, por ejemplo, los conceptos SOLTERO, MANZANA y ROJO– o, dicho de otro modo, son representaciones conceptuales codificadas y externalizadas mediante el lenguaje (Evans 2006, p. 494). Un aspecto interesante de los conceptos léxicos es que suele considerarse que las palabras heredan su significado de los conceptos que expresan.

específico de los conceptos léxicos, sino que aplicaría también a conceptos complejos no-léxicos.

La distinción entre conceptos primitivos y complejos se apoya directamente en la asunción del *principio de composicionalidad*, a saber, que el significado de los conceptos complejos es el resultado de la estructura y significado de sus conceptos constituyentes. La principal ventaja de asumir que los conceptos complejos son composicionales es que permite explicar tanto la *productividad* –capacidad de un sistema para producir / reconocer un número infinito de elementos distintos– como la *sistematicidad* –propiedad que tiene un sistema cuando produce/reconoce elementos con patrones definidos y predecibles– del pensamiento[6]. Así, por ejemplo, la lógica de predicados es productiva porque permite un número infinito de fórmulas bien formadas únicas, y es sistemática porque la producción/reconocimiento de la fórmula *eAd* (*Eva ama a David*) implica la capacidad de producir/reconocer la fórmula *dAe* (*David ama a Eva*).

Finalmente, aunque la diferencia entre conceptos primitivos y complejos habitualmente se presenta para el caso de sistemas representacionales, la posibilidad de distinguir entre ambas es independiente de cuál se asuma que es la naturaleza de los conceptos, siendo una distinción válida tanto si son representaciones mentales, habilidades o incluso entidades abstractas (Evans 1982; Zalta 2001).

1.2. Naturaleza de los conceptos

Una segunda cuestión clave que se plantea es la relativa al estatus ontológico de los conceptos. En este caso, la respuesta dada suele depender de cuál sea el área desde el que dicha respuesta se proporciona. Así, en el ámbito de la *psicología* –incluyendo aquí la neuropsicología, inteligencia artificial, filosofía de la mente y ciencia cognitiva– suele asumirse que los conceptos son entidades mentales particulares que representan a una cierta clase o categoría (Komatsu 1992; Laurence y Margolis 1999; Murphy 2002). Ahora bien, en *filosofía* el término “concepto” también es empleado de otros modos distintos, entre los que destaca su uso para referir, bien a la capacidad de poder tener actitudes proposicionales –creencias, deseos, etc.– con respecto a una cierta categoría (Machery 2009), bien a aquellas entidades abstractas con que en ocasiones se identifican los constituyentes de las proposiciones (Margolis y Laurence 2011a). Por consiguiente, cabe distinguir tres aproximaciones principales con respecto a qué son los conceptos, en función de si tales conceptos se conciben como *representaciones mentales*, *capacidades/habilidades* o *entidades abstractas*[7].

[6] La justificación habitualmente dada para la asunción del principio de composicionalidad en el ámbito de la mente es que si la productividad y sistematicidad del lenguaje se explican mediante el principio de composicionalidad, entonces esa misma debe ser la explicación para el caso de la productividad y sistematicidad del pensamiento (Fodor y Pylyshyn 1988; Fodor 1998, 2001). No obstante, esta tradicional justificación de la composicionalidad –en términos de su necesidad para explicar la productividad y sistematicidad del lenguaje y la cognición– ha sido recientemente puesta en tela de juicio por varios autores (Werning 2005; Pagin 2012).

[7] En ocasiones estos tres tipos de aproximaciones son referidas como subjetivista, cognitivista y objetivista (Glock 2010, p. 117). Los enfoques *subjetivistas* conciben a los conceptos como entidades o fenó-

1.2.1 Conceptos como representaciones mentales

Las aproximaciones subjetivistas a la noción de concepto, conforme a las cuales los conceptos son entidades mentales, se apoyan directamente en la teoría representacional de la mente (o TRM). Según la TRM, el pensamiento tiene lugar mediante un sistema de representación interno que, en su versión contemporánea[8], dispondría de una sintaxis (como la de un lenguaje) y una semántica composicional[9]. Esta concepción es habitualmente conocida como la *hipótesis del lenguaje del pensamiento* (Fodor 1975), y tiene como principal ventaja que, si las representaciones mentales se consideran composicionalmente estructuradas (esto es, con estructura interna composicional), entonces eso permite explicar la productividad y sistematicidad del pensamiento. Por todo ello, la visión de los conceptos como representaciones mentales es generalizada en el ámbito de la psicología, ciencia cognitiva y filosofía de la mente (Fodor 1998; Carruthers 2000; Millikan 2000; Prinz 2002; Margolis y Laurence 2007).

Así, bajo esta aproximación, el concepto SOLTERO es la representación mental –o entidad mental particular– de una determinada categoría que podría estar constituida por las representaciones de otros conceptos más simples (en el caso de que fuera una representación estructurada). No obstante, la aceptación de que los conceptos tienen estructura interna no obliga a identificar dichos conceptos con particulares mentales. O, en otras palabras, los conceptos bien podrían ser estructurados sin ser representaciones mentales[10]. En tal caso, una explicación alternativa sería que los conceptos son habilidades estructuradas de tipo psicológico o conductual –esto es, habilidades complejas constituidas por

menos psicológicos presentes en la mente de los sujetos (aquí estaría también incluida la visión de que los conceptos son aquellas entidades mentales que nos permiten tener actitudes proposicionales). Por el contrario, las concepciones *objetivistas* identifican los conceptos con entidades abstractas cuya existencia sería independiente de nuestras mentes. Finalmente, las aproximaciones *cognitivistas* –con su identificación de los conceptos con habilidades– están a medio camino entre el subjetivismo y el objetivismo pues, aún cuando aceptan que los conceptos tienen una dimensión mental, rechazan que puedan identificarse con particulares mentales.

[8] La visión contemporánea contrasta frente a la de los primeros defensores de la TRM (Locke 1690; Hume 1739), quienes consideraban que ese sistema de representación interno estaba basado en imágenes mentales a las que daban el nombre de *ideas* (Fodor 2003; Gauker 2011).

[9] Este punto es relevante, dado que la TRM explica la intencionalidad de los estados mentales (creencias, pensamientos, deseos, etc.) –esto es, el que dichos estados mentales sean sobre, o refieran a, algo– en términos de las propiedades semánticas de sus representaciones mentales asociadas.

[10] En este sentido va la doble crítica de Dennett (1977) a la aproximación representacional, según la cual: (a) Pueden tenerse actitudes proposicionales sobre algo sin disponer de una representación mental suya. Por ejemplo, la mayoría de las personas cree que las cebras no visten abrigos en la naturaleza, aún cuando nunca lo hayan considerado activamente con anterioridad. (b) Un sistema computacional puede "pensar" algo aún cuando no tenga una representación explícita de su contenido. Por ejemplo, un programa de ajedrez puede "pensar" que es bueno sacar a la reina pronto como algo emergente a su programación, y no en virtud de instanciar un conjunto de símbolos que expresen tal principio.

otras habilidades más simples[11]–, lo que conduce directamente a las aproximaciones cognitivistas[12].

Otra fuente de críticas a la aproximación representacional es por parte de aquellos autores que consideran que el conexionismo y/o la teoría de sistemas dinámicos son programas más prometedores que los enfoques computacionales clásicos (Rumelhart y McClelland 1986; Churchland 1989; Smolensky 1991; Van Gelder 1995; Elman *et al.* 1996; McClelland *et al.* 2010).

1.2.2 Conceptos como habilidades

Por su parte, las concepciones cognitivistas rechazan que los conceptos puedan identificarse con particulares mentales –por ejemplo, en el sentido de elementos de un lenguaje del pensamiento (Fodor 1975)–, aún cuando reconocen que tienen un carácter psicológico, y de ahí que los identifiquen con habilidades mentales (Geach 1957; Evans 1982; Dummett 1993; Kenny 2010).

Ahora bien, ¿qué habilidades mentales se requieren para que pueda decirse que alguien posee un determinado concepto? En este caso, las dos principales habilidades demandadas son las de reconocimiento e inferencia, esto es, se acepta que alguien posee un cierto concepto si es capaz de reconocer ejemplares de ese concepto, y si realiza inferencias específicas acerca de dichos ejemplares (que le conducen a reaccionar de manera distinta ante ellos que ante otras entidades del mundo). Así, por ejemplo, se diría que alguien dispone del concepto PERRO si tiene la capacidad para discriminar entre perros y no-perros, y además realiza inferencias específicas sobre los perros (que le hacen comportarse de un modo en particular cuando está ante ellos).

La principal motivación para los defensores de esta segunda aproximación a la naturaleza de los conceptos es su escepticismo sobre la existencia de las representaciones mentales, el cual se remota al menos al segundo Wittgenstein (1953). Ésa es la crítica de Searle (1992, pp. 212-214) a la teoría computacional de la mente, para quien la falacia del homúnculo es endémica en la ciencia cognitiva. Y también la de Dummett (1993, p. 98), cuando critica la posibilidad de explicar la comprensión de un primer lenguaje en términos de un segundo lenguaje, cuya comprensión (de este segundo lenguaje) demanda a su vez una explicación. El problema en último término es el de cómo se evita caer en un círculo vicioso cuando se emplea la noción de representación mental para explicar la significancia de los conceptos, en la medida en que para explicar la significancia del nuevo

[11] Tal posibilidad está en línea con las tesis de Evans (1982, p. 101), para quien los pensamientos podrían tener estructura en virtud de ser el resultado –complejo– de ejercer varias habilidades conceptuales (y no por estar compuestos de varios particulares mentales).

[12] Los enfoques cognitivistas –o visión de los conceptos como habilidades– son en ocasiones criticados porque, en ausencia de una teoría clara con respecto a lo que es una habilidad, podría incluso ser el caso que tales habilidades fuesen particulares mentales (Laurence y Margolis 1999, p. 6). Ahora bien, la noción de representación solo es más clara que la de habilidad en apariencia, pues la evaluación de las propiedades semánticas –o estructuras portadoras de información– asociadas a una cierta representación requiere de un intérprete (esto es, de un sistema que interprete dicha información o propiedades semánticas). El problema es que cuando dicho intérprete es incluido en la ecuación, la noción de representación mental se aproxima peligrosamente a la noción de habilidad mental.

nivel introducido –representaciones mentales– se requeriría de otro nivel de representación (y así sucesivamente). Este tipo de críticas apuntan a lo que Crane (1995) llama el puzle de la representación, esto es, al problema de cómo es posible que algo sirva para representar a alguna otra cosa.

Por su parte, el problema de la noción de habilidad –entendida en términos de discriminación e inferencia– es su excesiva vaguedad, e insuficiente especificación de cómo funcionan los procesos psicológicos subyacentes. De hecho, sin una mayor concreción de cómo están articuladas esas dos capacidades, la visión de los conceptos como habilidad podría consistir en una aproximación basada en representaciones mentales[13], siendo ésta una de las dos líneas de crítica recibidas desde el ámbito representacionalista. La otra es que, si los enfoques cognitivistas rechazasen explícitamente la existencia de representaciones mentales, entonces dichos enfoques tendrían problemas para explicar la productividad del pensamiento[14].

1.2.3 Conceptos como entidades abstractas

Esta tercera aproximación a la naturaleza de los conceptos podría ser descrita como objetivista, en la medida en que considera que los conceptos son aquellas entidades abstractas que constituyen las proposiciones (y, por lo tanto, ontológicamente objetivas[15]), susceptibles de ser identificadas con los sentidos –o modos de presentación– postulados por Frege (Bealer 1982; Peacocke 1992; Zalta 2001). Los defensores de este enfoque consideran –en línea con Frege (1892)– que los conceptos (sentidos o modos de presentación) median entre los nombres (o signos) de los objetos y sus referentes[16].

El *sentido* de una expresión lingüística es su contenido cognitivo y, por ello, aquello que permite determinar cuál es su referente, siendo la comprensión de dicho sentido lo que da acceso al referente. En esta concepción distintos términos pueden referir a un mismo objeto (haciéndolo con modos de presentación diferentes[17]). Así, por ejemplo, las

[13] Obsérvese que los enfoques cognitivistas no son simplemente incompatibles con la existencia de representaciones mentales –o particulares mentales–, sino con la posibilidad de identificar dichos particulares mentales con los conceptos.

[14] Ahora bien, esta opinión puede no ser compartida desde el ámbito cognitivista, desde donde podría argumentarse que los conceptos complejos son habilidades complejas constituidas por otras habilidades conceptuales más simples (Evans 1982).

[15] Frente a la naturaleza más o menos subjetiva de las otras dos aproximaciones, que de una u otra forma aceptaban la tesis de que los conceptos "están en la mente". Margolis y Laurence (2007) refieren a este enfoque como *visión semántica* de los conceptos.

[16] Esto, expresado en términos del triángulo semiótico (Ogden y Richards 1923), se correspondería con la mediación de la mente –o pensamiento– entre el lenguaje y el mundo, en el seno de la explicación de cómo los sujetos (mentes) son capaces de referir a los objetos del mundo (referentes) por medio del lenguaje (nombres o signos).

[17] En palabras de Peacocke (1992, p. 2), dos conceptos son distintos –o, en la terminología de Frege, tienen modos de presentación distintos– si y solo si (a) existen dos contenidos proposicionales completos que difieren entre sí solamente en que uno de ellos contiene a uno de esos conceptos substituido por el otro (en uno o varios lugares), y (b) uno de esos contenidos proposicionales es informativo, mientras que el otro no lo es.

expresiones “dos más cinco” y “tres más cuatro” tendrían el mismo referente (número siete), aún cuando sus modos de presentación (sentidos) y signos son distintos. Y, análogamente, las expresiones “el autor de Waverley” y “Walter Scott” tienen el mismo referente, aunque lo expresan por medio de sentidos y signos lingüísticos diferentes. En todos estos casos, son las diferencias en el modo de presentación las que determinan los distintos contenidos cognitivos (o conceptos) expresados.

La motivación de los defensores de esta tercera alternativa es diferente a la de los partidarios de la visión cognitivista. Aún cuando ambos se distancian de la aproximación representacional, unos y otros lo hacen por razones diferentes. En este caso, el principal problema que los objetivistas ven en la noción de representación mental es su carácter subjetivo, lo que impediría la existencia de una comunicación exitosa (si, como es asumido por ellos, la comunicación precisa de significados compartidos que la soporten, y no de representaciones mentales dependientes de la experiencia subjetiva de cada uno). Obviamente, los enfoques cognitivistas adolecerían de ese mismo problema, debido al carácter subjetivo de las habilidades mentales asumidas por ellos.

1.3. Origen de los conceptos: nativismo vs empirismo

Otra cuestión que habitualmente se plantea con respecto a los conceptos –y, posiblemente, una de las más antiguas en el tiempo– es la relativa a cuál es el origen de dichos conceptos, en el sentido de si son innatamente heredados o aprendidos a partir de la experiencia. Esta cuestión conecta directamente con el problema de si los conceptos pueden ser adquiridos y, de ser así, cómo puede tener lugar dicha adquisición.

1.3.1 El debate empirista-nativista

La cuestión de dónde vienen los conceptos no solo es uno de los grandes desafíos a los que se enfrenta la ciencia cognitiva actual, sino que también fue uno de los temas críticos a los que se enfrentaron la filosofía antigua y moderna. En este caso, dos han sido las principales respuestas que ha recibido la cuestión sobre el origen de los conceptos: (a) *nativista*, como tradición que se extiende desde Platón hasta los nativistas contemporáneos (tales como Fodor y Carey), pasando por Kant y los filósofos racionalistas modernos, conforme a la cual muchos conceptos son innatos; (b) *empirista*, como tradición que va desde Aristóteles hasta el empirismo moderno y contemporáneo, y que mantiene que los conceptos proceden de la acumulación de experiencia sensorial, y que muy pocos, si alguno, de ellos son innatos[18].

[18] No obstante, cabría argumentar –conforme hacen Margolis y Laurence (2013)– que tanto empiristas como nativistas aceptan la existencia de sistemas de aprendizaje innatos, razón por la cual el nativismo no puede ser definido como la defensa de la tesis de que muchos conceptos son innatos (en el sentido de ser adquiridos mediante una base psicológica innata), puesto que el empirismo también podría cumplir esa condición. Y, por otro lado, dado que tanto empiristas como nativistas pueden estar dispuestos a aceptar que (algunos) conceptos pueden ser aprendidos de la experiencia –junto con una cierta base de adquisición–, el empirismo no puede ser entonces definido por la tesis de que los conceptos proceden de la acumulación de experiencia sensorial, puesto que el nativismo también podría aceptar esa condición. Por todo ello, el debate empirista-nativista contemporáneo no debería girar en torno

Aunque nativismo y empirismo fueron durante siglos las dos principales aproximaciones en pugna por explicar cómo los conceptos se adquieren, a mediados del siglo veinte –y tras siglos de discusión– el debate entre ellas parecía haber llegado a su término, con el empirismo como opción vencedora. No obstante, la discusión resurgió en la segunda mitad del siglo con los trabajos de Chomsky (1959; 1965), principalmente con su argumento de la pobreza del estímulo y su gramática generativa. Esos nuevos argumentos y evidencias a favor del nativismo dejaron la pelota en el tejado empirista, desde donde se debía una explicación al respecto. No obstante, los problemas del empirismo no terminaban allí, dado que poco después tuvo que hacer frente a los argumentos de Fodor a favor del nativismo y en contra del empirismo, y a su tesis de que los conceptos carecen de estructura interna (Fodor 1975; 1981a). Todo ello dio lugar a una reacción desde el lado empirista, con la aparición de nuevas propuestas que intentaban dar cuenta de los problemas y críticas esgrimidos en contra de la teoría empirista clásica. Esos esfuerzos cristalizaron en las tres aproximaciones actualmente más populares dentro del empirismo, a saber: *conexionismo*, *teoría de sistemas dinámicos* y *bayesianismo*. En el presente trabajo no profundizaré en los detalles de ninguna de estas aproximaciones, sino que me centraré en otra perspectiva –no dominante– dentro del empirismo. Tal enfoque consistirá en una caracterización del marco empirista que explique la adquisición de conceptos mediante un proceso iterativo, constituido por la ejecución secuencial de una reducción dimensional seguida de un proceso de reconocimiento de patrones.

1.3.2 Argumentos nativistas

Una de las primeras razones positivas dadas a favor del nativismo en el siglo pasado procedía de las investigaciones relativas a la adquisición del lenguaje natural, un ámbito en el que Chomsky (1959; 1967; 1975; 1980) consideraba que la gramática de un lenguaje natural no puede adquirirse a partir de los limitados datos de que disponen los niños que lo aprenden[19], en lo que se conoce como el *argumento de la pobreza del estímulo*. Su conclusión era que la adquisición del lenguaje está basada en un conjunto de disposiciones (asentadas sobre unos principios innatos a los que Chomsky llama *gramática universal*) que constriñen el modo en que el lenguaje es aprendido[20].

Sobre esta base, una sencilla reformulación del argumento permite su aplicación al caso de la mente y la adquisición de conceptos. En este ámbito, la idea es que la información de que disponen los sujetos no es suficiente para explicar cómo se adquieren los conceptos solo por medio de sistemas de aprendizaje de propósito general. Aquí, la tesis nativista es que la diferencia entre la entrada experiencial del sujeto y la información mínima nece-

al carácter aprendido o innato de los conceptos, sino en torno al carácter de los mecanismos cognitivos que subyacen al proceso de adquisición de dichos conceptos. Bajo ese prisma, el empirismo se caracterizaría por estar soportado por mecanismos de propósito general, mientras que lo característico del nativismo es que estaría soportado por mecanismos específicos de dominio (Pinker 2002; Spelke y Kinzler 2007).

[19] La mayoría de los argumentos nativistas son, en el fondo, argumentos en contra del empirismo que, por exclusión, contribuyen a la afirmación de posturas de tipo innatista.

[20] Líneas de argumentación semejantes pueden encontrarse en Goodman (1967) y Putnam (1967).

saria –para adquirir un cierto concepto o habilidad– es provista por sistemas de aprendizaje innatos y específicos de dominio.

Ahora bien, el argumento de la pobreza del estímulo ha sido objeto de dos tipos de crítica. Por un lado, ciertos filósofos sostienen que no existen evidencias suficientes a favor de un entorno tan pobre como el asumido por los nativistas (Cowie 1999). Por otro lado, hay autores que consideran que la potencia de los sistemas de aprendizaje de propósito general ha sido subestimada, y que ese tipo de sistemas en realidad sí podría permitir la adquisición de conceptos generales (Prinz 2002)[21]. Ambas críticas han sido a su vez cuestionadas desde el ámbito nativista (Laurence y Margolis 2001, 2015; Margolis y Laurence 2013), y éstas a su vez respondidas por partidarios del empirismo, en una dinámica que deja a este argumento en una situación de impasse.

Un segundo argumento contrario al empirismo –y en favor del nativismo– es el *argumento de los animales*, según el cual la existencia de un gran número de sistemas de aprendizaje específicos en el reino animal[22] (algunos de ellos compartidos entre especies) sugiere que la mente humana podría estar constituida por sistemas de aprendizaje especializados e innatos (Margolis y Laurence 2013, p. 702). En este caso el problema es que existen otras explicaciones alternativas que también darían cuenta de cómo esas habilidades se adquieren, las cuales (aún si fueran desempeñadas por sistemas de propósito específico) podrían no ser innatas, sino el producto de otros sistemas de aprendizaje de propósito general[23]. Esta posibilidad enlaza con las críticas al argumento de la pobreza del estímulo anteriormente mencionadas, pues el entorno podría no ser tan pobre como se supone y, además, la potencia de los sistemas de aprendizaje de propósito general podría estar siendo subestimada. O, dicho de otro modo, las críticas en contra del argumento de la pobreza del estímulo son también críticas en contra del argumento de los animales, mediadas por la hipótesis de que los subsistemas –específicos de dominio– responsables del aprendizaje hubiesen sido previamente adquiridos por medio de sistemas de propósito general.

No obstante, posiblemente una de las objeciones más serias al empirismo (y, por tanto, razón para abrazar las tesis innatistas) son los *argumentos de Fodor* a favor del nativismo y en contra del empirismo (Fodor 1975; 1981a). Conforme a ellos, si los conceptos se adquieren mediante un proceso de formación y comprobación de hipótesis –y, en opinión de Fodor, ése es el único método de aprendizaje posible–, entonces los conceptos no

[21] No obstante, estas dos críticas podrían ser consideradas de modo conjunto, puesto que (i) un entorno informacional más rico podría estar disponible para sujetos cuyos sistemas de aprendizaje de propósito general fuesen más potentes que lo asumido por los nativistas; y (ii) si la potencia de los sistemas de adquisición hubiera sido subestimada, tal vez dichos sistemas podrían hacer uso de un mayor número de claves informacionales presentes en el entorno.

[22] Sistemas, por ejemplo, para elaborar mapas mentales del entorno, para elegir lugares en donde buscar alimento –en función de la tasa de retorno esperado–, para evitar comida venenosa, para señalar la presencia de predadores, para detectar familiares o identificar las relaciones en una jerarquía social, etc.

[23] Esto está en línea con la sugerencia de Goodman (1967) de que los subsistemas especializados responsables de la adquisición del lenguaje podrían estar a su vez soportados por otras habilidades generales previamente adquiridas.

se pueden aprender porque sus constituyentes deberían de estar disponibles con anterioridad al proceso de aprendizaje (con objeto de poder formular la hipótesis sobre ese concepto que será testeada), lo que da lugar a un regreso al infinito cuando se intenta explicar cómo se adquieren los constituyentes más básicos de los conceptos (esto es, los llamados *conceptos primitivos*). El argumento de Fodor es, por consiguiente, que los modelos empiristas caen en circularidad cuando intentan explicar cómo se adquieren los conceptos primitivos los cuales, en consecuencia, deberían ser innatos.

Cuando, como hace Fodor (1981a), se considera que los restantes conceptos (*complejos*) resultan de la clausura del conjunto de conceptos primitivos (e innatos) bajo mecanismos combinatorios (también considerados innatos)[24], entonces resulta fácil derivar la tesis de que todos los conceptos están innatamente determinados. Y, aunque se han dado múltiples respuestas a los argumentos de Fodor[25], casi todas resultan inadecuadas por uno u otro motivo, no siendo ninguna plenamente satisfactoria (*cfr*. Laurence y Margolis 2002).

1.3.3 Argumentos empiristas

En cuanto a los argumentos en defensa del empirismo[26], la mayoría de ellos están presentes en mayor o menor medida en la argumentación de Locke (1690) en favor de esta postura, razón por la cual la primera parte de esta sección estará dedicada a presentar una revisión actualizada de los argumentos lockeanos más importantes[27]. Aunque en Locke encontramos al menos cinco argumentos en favor del empirismo, los más relevantes desde un punto de vista contemporáneo están asociados a las dos tesis siguientes: (i) no hay ideas –o conceptos– universalmente aceptadas / mantenidas; y (ii) las ideas innatas son innecesarias.

— *No hay conceptos universalmente aceptados, por lo que ninguno de ellos es innato:* este argumento daría cuenta de por qué las personas que carecen de cierto tipo de sensaciones también carecen de los correspondientes conceptos. Esto explicaría por qué las personas ciegas no tienen conceptos de colores, y las personas sordas no tienen conceptos de sonidos (Hume 1741, II 7). Carruthers (1992, p. 49) denomina a esta crítica como *argumento de los niños*, dado que si existiesen conceptos o

[24] O, en otras palabras, si se acepta que la potencia expresiva de nuestro sistema conceptual se encuentra determinada por sus conceptos primitivos y los principios combinatorios que los gobiernan.

[25] Para un repaso de las diferentes respuestas dadas a los planteamientos de Fodor véanse Laurence y Margolis (2002) y Carey (2015).

[26] Y, por exclusión, contrarios al nativismo (de modo análogo a como ocurría en el caso de los argumentos nativistas, que también eran argumentos en contra del empirismo).

[27] Indico que será una *revisión actualizada* porque, en algunos casos, lo que se presentará no será el argumento literal de Locke, sino una revisión del mismo que –manteniendo aquella parte válida del argumento– responda a las posibles críticas de que podría ser objeto el argumento original. Y lo será, no para todos sus argumentos, sino tan solo para los *argumentos más relevantes*, en el sentido de que –bien en su forma original, bien en su forma revisada–, continúen siendo argumentos válidos desde una perspectiva contemporánea.

habilidades innatas entonces deberían estar presentes en todo el mundo, teniendo que estar disponibles a la conciencia desde el nacimiento.

Una respuesta habitual a este argumento es que los conceptos innatos podrían estar latentes en los niños[28], en la medida en que sabemos que existe conocimiento que no está accesible a la conciencia pero sí presente en la memoria. Ejemplos de esto serían las pérdidas temporales de memorias, en las que en un cierto momento no se recuerda un hecho aún cuando tal conocimiento sí está en la memoria (en la medida en que sea recordado en un momento posterior).

Otra respuesta es que los conceptos innatos podrían estar latentes en el sentido de estar determinados a aparecer de manera innata en un cierto momento de su desarrollo cognitivo normal, sin importar cuál fuera su biografía, formación o experiencia. Una concepción así tendría que aceptar que una cierta cantidad de experiencia es necesaria para la operación normal de la mente y que, con ello, los conceptos innatos aparezcan; pero sin que una experiencia concreta fuera crítica para adquirir un concepto dado[29].

— *Los conceptos innatos son innecesarios, pues todos nuestros conceptos se pueden explicar como derivados a partir de la experiencia:* en efecto, si el origen de todo concepto pudiera explicarse recurriendo únicamente a la experiencia, entonces los conceptos innatos serían innecesarios por motivos de parsimonia[30].

La respuesta tradicional a este segundo argumento ha sido que el antecedente carece de base, en la medida en que el empirismo nunca ha sido capaz de proporcionar una explicación adecuada de cómo todos nuestros conceptos pueden ser derivados a partir de la experiencia[31]. No obstante, esta respuesta es menos concluyente cuando se define el empirismo sobre la base del tipo de sistemas de aprendizaje que lo soportan –esto es, sistemas de propósito general– (Margolis and Laurence 2013).

Sin embargo, posiblemente el principal problema de este argumento empirista es el de circularidad. En este caso, la tesis lockeana de que derivamos los conceptos simples de la experiencia por abstracción –bajo la asunción de que podemos aislar aquellos rasgos comunes a esos conceptos– resulta circular, pues para saber que unos objetos tienen un rasgo en común es preciso disponer antes del concepto de

[28] Esta respuesta ya está presente en la réplica de Leibniz a Locke en sus *Nuevos ensayos sobre el entendimiento humano*, según la cual los conceptos innatos podrían existir en la mente como disposiciones, actitudes o pre-formaciones hacia el desarrollo y comprensión de determinados pensamientos (Leibniz 1765, I i 11).

[29] Esto es, bastaría con que la experiencia fuera lo suficientemente rica y variada para llegar a la formación de ese concepto, sin requerir una correspondencia específica entre experiencias y conceptos adquiridos. Carruthers (1992, p. 51) llama a esto *hipótesis del desencadenamiento general del conocimiento innato*.

[30] Obsérvese que el antecedente de este argumento constituye la tesis principal –y definitoria– del empirismo, según la cual todo concepto proviene de nuestra experiencia.

[31] Alternativamente, si el empirista demandase aceptar el antecedente como hipótesis estaría incurriendo en petición de principio pues, como se ha indicado, dicho antecedente no es otra cosa que la definición de empirismo.

> ese rasgo. Es decir, si los conceptos se adquieren por formación y comprobación de hipótesis, entonces se requiere poseer de antemano los conceptos que aparecen en la hipótesis, siendo Fodor (1975; 1980a; 2008) uno de los principales defensores contemporáneos de esta crítica al empirismo.

Un tercer argumento menor en favor del empirismo es que toda aproximación nativista deriva en nativismo radical –tal y como sostiene Fodor (1981a)[32]–, siendo ésta una tesis que muy pocos estarían dispuestos a considerar como aceptable, en la medida en que resulta incompatible con la evolución (Putnam 1988, p. 15).

1.4. Grado de dependencia contextual: invariantismo vs contextualismo

En cuanto a los diferentes enfoques que pueden adoptarse acerca del grado de dependencia contextual que cabe atribuir a los conceptos, las dos principales –y opuestas– posturas son la invariantista y la contextualista. La principal diferencia entre estas dos aproximaciones es que, mientras que para el *invariantismo* los conceptos son cuerpos de conocimiento estables entre individuos y tiempos (Keil 1994; Fodor 1998; Mazzone y Lalumera 2010; Barsalou 2012; Machery 2015), el *contextualismo* los identifica con constructos creados de modo específico para cada ocasión (Barsalou 1987, 1992; Sperber y Wilson 1995; Carston 2002; Prinz 2002; Hoenig *et al.* 2008; Malt 2010; Kiefer y Pulvermüller 2012; Casasanto y Lupyan 2015; Lebois, Wilson-Mendenhall y Barsalou 2015; Yee y Thompson-Schill 2016)[33].

La razón por la que existen dos aproximaciones a la noción de concepto tan distintas como las anteriores es que cada una de ellas da cuenta de un fenómeno cognitivo clave, cuya explicación resulta crucial para toda teoría sobre los conceptos[34]. Por un lado, los cuerpos de conocimiento estables –asumidos por el invariantismo– permiten explicar cómo es posible que acumulemos nuevo conocimiento acerca de una categoría. Por otro lado, la dependencia del contexto de los constructos específicos-de-uso –asumidos por el contextualismo– permite dar cuenta de nuestra capacidad para adaptarnos ante entornos cambiantes.

[32] Para una presentación y discusión tanto de la postura Fodor, como de sus posibles críticas, véase Laurence y Margolis (2002).

[33] En cierto modo puede decirse que el contextualismo sistematiza lo que en mayor o menor medida ya estaba presente en el *principio de contexto* de Frege –el cual recomendaba "nunca preguntar por el significado de una palabra aislada, sino solo en el contexto de una proposición" (Frege 1884, p. xxii)–, y que más adelante recogerá Wittgenstein en su *Tractatus* –cuando indique que "solo en el contexto de una proposición tiene el nombre significado" y "una expresión solo tiene significado en una proposición" (Wittgenstein 1922, §3.3 y §3.314)–.

[34] No obstante, las dos posturas aquí presentadas se corresponden con los dos extremos un espectro mucho más amplio en el que, por lo general, ambos –contextualistas e invariantistas– asumen: (i) que cierta información almacenada sobre una categoría se activa siempre, independientemente de cuál sea el contexto; y (ii) que en ocasiones se precisa de información dependiente del contexto para explicar el comportamiento en circunstancias no habituales (Löhr 2017). Bajo este prisma, la discrepancia surge con respecto a si la información siempre activada mencionada en el punto (i) es suficiente, o no, como para explicar nuestro comportamiento en las situaciones normales.

1.4.1 Enfoque invariantista

La visión tradicional –o enfoque *invariantista*– identifica los conceptos con cuerpos de conocimiento estables entre individuos y tiempos. Sobre esta base, los partidarios del invariantismo consideran que un concepto es aquel conocimiento (sobre una cierta categoría) que nuestro sistema cognitivo siempre recupera independientemente de cuál sea el contexto. Esta concepción explica de manera inmediata la estabilidad del pensamiento y la comunicación, a nivel tanto intrapersonal como interpersonal. O, en otras palabras, este enfoque da cuenta con facilidad tanto de la acumulación de conocimiento por parte de los sujetos, como de la capacidad de éstos para comunicarse entre ellos. Las principales razones dadas en su favor suelen ser las siguientes:

— *Si los conceptos no fuesen estables para un mismo sujeto S*, entonces no habría nada que proporcionase la continuidad que un concepto *C* necesita para acumular nueva información sobre él. Esto es, el sujeto *S* no podría acumular nuevo conocimiento sobre el concepto *C* pues no habría modo de reconocer nuevas instancias de *C* en distintos momentos de tiempo.

— *Si los conceptos no fuesen estables y compartidos entre los interlocutores de una conversación*, entonces la mutua comprensión de los mensajes intercambiados no sería posible[35]. El motivo de ello es que, aunque el hablante significase *C* con su proferencia del término *t*, podría ocurrir que el oyente interpretase *t* como otro concepto *C'*, distinto de *C*.

No obstante, el principal problema del invariantismo es el de explicar cómo es posible la invariancia de los conceptos, lo que en ocasiones es referido como el *problema de la estabilidad conceptual*. Este problema presenta dos vertientes, cada una de ellas asociada a una de las dos razones anteriores. Por un lado estaría la cuestión de cómo el contenido de un concepto puede permanecer invariante entre cambios de creencias. Por el otro está el problema de cómo puede ser que personas con creencias diferentes posean conceptos con idénticos –o similares– contenidos[36,37].

1.4.2 Enfoque contextualista

Frente al invariantismo, otros autores sostienen que muchos conceptos dependen del contexto, en el sentido de que serían constructos creados al vuelo de manera específica

[35] Esta condición deseada para los conceptos –como requisito para la comunicación– es también referida como el *desiderátum de publicidad* (Prinz 2002).

[36] O, si se prefiere no hablar del "contenido" de un concepto, este doble problema podría parafrasearse en los términos siguientes: ¿Cómo un concepto puede permanecer invariante entre cambios de creencias? ¿Cómo personas con creencias distintas pueden poseer los mismos –o similares– conceptos?

[37] Este segundo problema se encuentra íntimamente relacionado con el *problema del significado compartido* (en el ámbito del lenguaje), y con el *problema del acuerdo/desacuerdo* (en el ámbito de la epistemología), relativos –respectivamente– a cómo es posible que personas con biografías, experiencias y creencias distintas compartan los mismos significados y puedan llegar a acuerdos/desacuerdos.

para cada ocasión[38]. Dicho esto, no hay un único modo en que el contextualismo puede concebirse. Por un lado estaría la corriente dominante, según la cual la información activada / recuperada / accedida sobre un concepto siempre depende del contexto (Barsalou 1987; Casasanto y Lupyan 2015). No obstante, en ocasiones se sostiene incluso que la información almacenada en los conceptos cambia constantemente con el contexto (Löhr 2017). En todo caso, todas estas diversas posturas consideran –de uno u otro modo– que los conceptos están indisolublemente ligados al contexto en el que aparecen, razón por la cual sería imposible distinguir completa y claramente entre un concepto y su contexto.

Los contextualistas respaldan estas afirmaciones en evidencias empíricas que muestran que los conceptos dependen de nuestra experiencia –tanto de corto como de largo plazo–, de la situación, de los objetivos del sujeto, del paso del tiempo, etc. (Barclay *et al.* 1974; Roth y Shoben 1983; Barsalou 1987, 1993; Yee y Thompson-Schill 2016)[39]. En esta misma línea, la ventaja principal de la concepción *contextualista* es que explica la adaptación de nuestro comportamiento ante contextos / entornos / circunstancias cambiantes (lo cual sería difícil de explicar si los conceptos fueran invariantes)[40].

En cuanto a los problemas a los que se enfrenta contextualismo, es muy posible que el mayor de ellos sea el hecho de que prácticamente nadie proporciona una definición de lo que es el contexto, ni de cómo podría estar operacionalmente articulado (Bloch-Mullins 2015), lo que hace que el contextualismo tan siquiera esté completamente especificado. En este caso la dificultad estriba en que es necesaria una definición (del contexto) en la que cada nueva situación se corresponda con un nuevo contexto y, al tiempo, resulte lo bastante potente como para explicar todos los distintos fenómenos de dependencia contextual observados[41].

* * *

Finalmente, aún a pesar de los argumentos y evidencias esgrimidos por una y otra parte para confirmar o rechazar la posibilidad de que pueda haber cuerpos de conocimiento estables sobre las categorías, ninguna de ellas ha sido crucial para decidir el debate, por lo que a día de hoy sigue sin existir un consenso al respecto.

[38] Barclay *et al.* (1974) comprobaron que los rasgos relevantes del concepto PIANO dependen del contexto: en un contexto de *producir música* lo serán sus propiedades musicales, mientras que en uno de *mover mobiliario* lo será su peso. Similarmente Barsalou (1993) muestra que al considerar el concepto PERIÓDICO en un contexto habitual, *inflamable* no es uno de sus rasgos asociados, pero sí que lo es en un contexto de *hacer fuego*. Por otro lado, estudios recientes basados en imágenes por resonancia magnética funcional (IRMf) y potenciales relacionados con eventos (ERPs) confirmarían el carácter flexible de los conceptos, y su constitución por elementos modales obtenidos en función del contexto (Hoenig *et al.* 2008; Kiefer y Pullvermüller 2012). Para una revisión de la creciente evidencia empírica en favor de que el procesamiento de los conceptos depende del contexto véase Yee y Thompson-Schill (2016).

[39] Este tipo de evidencias son, por lo general, investigaciones realizadas mediante el estudio de tareas / comportamientos cognitivos de alto-nivel –tales como reconocimiento de objetos y realización de inferencias–, los cuales resultan ser altamente dependientes del contexto.

[40] En ocasiones el *invariantismo* es puesto en correspondencia con la visión amodal de los conceptos, mientras que al *contextualismo* se lo identifica con la *embodied cognition* (Kiefer y Pullvermüller 2012; Bloch-Mullins 2015), lo que refuerza el contraste entre ambas posturas.

[41] Obviamente, la caracterización del contexto queda fuera del alcance de la presente tesis doctoral.

1.5. Relaciones cruzadas entre debates: ¿qué alternativas son viables?

Ahora bien, los dos debates anteriores –a saber, nativismo frente a empirismo, e invariantismo frente a contextualismo– no son independientes entre sí. De hecho, el propósito de la presente sección es revisar las relaciones existentes entre esas cuatro posturas, y determinar qué combinaciones son teóricamente viables[42].

Para ello, únicamente consideraré las dos posturas extremas en ambos debates, sin tener en cuenta otras posiciones intermedias que podrían adoptarse[43]. Esto es, asumiré que empirismo y nativismo son aproximaciones disjuntas y que cubren al completo el dominio de explicaciones posibles sobre el origen de los conceptos; y que invariantismo y contextualismo también son aproximaciones disjuntas y que cubren al completo el dominio de grados de dependencia contextual que pueden mostrar los conceptos. Estas dos asunciones pueden expresarse formalmente mediante las siguientes relaciones de equivalencia:

$$\text{Empirismo} \leftrightarrow \neg\, \text{Nativismo}$$

$$\text{Invariantismo} \leftrightarrow \neg\, \text{Contextualismo}$$

1.5.1 Invariantismo implica nativismo

La primera relación que puede establecerse entre el grado de dependencia contextual de los conceptos –en el lugar del antecedente– y el origen de tales conceptos –en el consecuente–, es la relación de implicación entre invariantismo y nativismo, esto es:

$$\text{Invariantismo} \rightarrow \text{Nativismo}$$

En este caso, el argumento en favor de la implicación anterior –al que daré el nombre de ARGUMENTO IiN– podría articularse en los términos siguientes:

(1) Los conceptos son invariantes entre individuos y tiempos. (*Invariantismo*)
(2) Para cada categoría, el mismo concepto es compartido por todo individuo.
(3) Cada individuo tiene una biografía y experiencia distintas, obtenidas a partir de diferentes entornos y contextos.
(4) Aceptando (3), resulta altamente implausible que distintos individuos –expuestos a experiencias y ambientes diferentes– hayan aprendido el mismo e idéntico concepto a partir de una pluralidad de experiencias distintas de una categoría.
(5) La explicación más parsimoniosa es que los sujetos comparten el mismo concepto de una cierta categoría porque dicho concepto es idénticamente heredado por todos ellos.
(6) Por consiguiente, todos los conceptos serían innatos. (*Nativismo*)

[42] La Tabla 1.1 presenta resumidamente cuáles son esas alternativas teóricamente viables.

[43] Con respecto a esto, y aún cuando existirían argumentos tanto a favor como en contra de una asunción como la anterior –así, por ejemplo, Fodor (1981a) sostiene que toda postura innatista deriva en innatismo radical, y Cappelen y Lepore (2005) defienden que todo contextualista moderado tendría que adscribir las tesis del contextualismo radical–, dejaré la discusión de tales argumentos fuera del ámbito del presente trabajo.

Por lo tanto, la conclusión es que si el concepto de una determinada categoría es invariante para todo individuo, entonces también debe ser innato.

Obsérvese que lo que ha permitido formular un argumento como el anterior ha sido una inferencia hacia la mejor explicación realizada en base a la experiencia de comprensión mutua. De hecho, la motivación que subyace al invariantismo se podría resumir de manera esquemática del modo siguiente:

[1] Los sujetos son capaces de acumular conocimiento sobre una categoría.
[2] Los sujetos son capaces de comunicarse exitosamente[44].
[3] Ni [1] ni [2] son posibles si los conceptos no fueran invariantes[45].
[4] Por inferencia hacia la mejor explicación, los conceptos son invariantes.

No obstante, este último esquema de argumentación –puntos [1] a [4]– no pretende ser un argumento a favor del invariantismo, sino tan solo un esbozo que muestre una de las posibles motivaciones que puede guiar al invariantista cuando éste asume la tesis de que *los conceptos son invariantes entre individuos y tiempos* (esto es, la premisa (1) del ARGUMENTO IiN). Por ello, las posibles críticas que pudieran realizarse a este esquema de motivación no supondrían un problema para el ARGUMENTO IiN, el cual funciona igualmente tomando simplemente como punto de partida la premisa (1) –como tesis principal asumida por el invariantista–.

1.5.2 Empirismo implica contextualismo

Por otro lado, cuando se pone el origen de los conceptos en el lado del antecedente y el grado de dependencia contextual en el del consecuente, lo que encontramos es una relación de implicación entre empirismo y contextualismo, esto es:

$$\text{Empirismo} \rightarrow \text{Contextualismo}$$

El argumento en favor de esta implicación –al que llamaré ARGUMENTO EiC– se podría articular como sigue:

(1) Los conceptos son adquiridos a partir de la experiencia. (*Empirismo*)
(2) Para cada categoría, su concepto será aprendido por los distintos individuos sobre la base de las experiencias y ambientes a que hayan estado expuestos en sus biografías.
(3) Pero las biografías de individuos distintos son diferentes, por lo que también lo serán las experiencias que hayan tenido con respecto a esa categoría[46].

[44] Claramente, el punto [2] supone un compromiso del invariantista con la intersubjetividad. No obstante, éste no es un compromiso inesperado, en la medida en que si el invariantismo no explicara cómo es posible la comunicación, perdería entonces una de sus dos principales razones de ser.

[45] Esta premisa, no obstante, puede ser cuestionada –tal y como se hace de modo habitual desde el ámbito contextualista–.

[46] Con respecto a este punto (3) y el siguiente punto (4) cabría replicar que es necesario asegurar [i] que las biografías de los sujetos son suficiente y relevantemente diferentes, y [ii] que no hay mecanismos co-

(4) Sobre la base de (3), resulta implausible que distintos individuos –con biografías y experiencias diferentes– hayan adquirido el mismo concepto de una categoría sobre la base de una pluralidad de experiencias –potencialmente muy distintas–[47].

(5) La explicación más parsimoniosa es que los conceptos no son invariantes entre individuos[48], sino que dependen –al menos– de cuáles hayan sido la biografía y experiencias de tales sujetos.

(6) En consecuencia, todo concepto dependería del contexto. (*Contextualismo*)

Aquí la conclusión es que, si el concepto de una categoría resulta de un proceso de aprendizaje basado en la experiencia, entonces ese concepto no es invariante para todo individuo, por lo que depende del contexto.

En este caso el punto de partida del ARGUMENTO EiC es la tesis básica del empirismo –o premisa (1) del argumento–, cuya motivación podría estructurarse como inferencia hacia la mejor explicación, en los términos siguientes[49]:

[1] Los seres humanos (como especie) han cambiado a lo largo del tiempo.

[2] El entorno en el que los seres humanos han vivido también ha sufrido cambios.

[3] Si las categorías en el mundo han cambiado, ¿cómo es posible que sus conceptos asociados fueran innatos?[50]

rrectores (por ejemplo, el uso de un lenguaje común) que –aún para biografías distintas– permitan alinear las categorías consideradas.

En relación con esta objeción, podría decirse que la lectura que Kripke (1982) hace de la paradoja wittgensteiniana de seguir una regla (Wittgenstein 1953: §201) constituye: [a] un argumento en contra de la posibilidad de que tales mecanismos correctores puedan existir (en la medida en que el lenguaje no permitiría dicha alineación de categorías pues, en último término, todo lenguaje es privado); y, por lo tanto, [b] un argumento en favor de que toda diferencia biográfica es una potencial fuente de desigualdad entre los conceptos adquiridos. (Obsérvese que la lectura de Kripke está en línea con las críticas de Fodor y Lepore (1992) a la visión conexionista, debidas a la imposibilidad de estar seguros de que los espacios de activación ocultos asociados a un cierto concepto son los mismos para los distintos individuos, lo cual no es un problema específico del enfoque conexionista sino un fenómeno general susceptible de alcanzar a todos los enfoques –como incapacidad para explicar cómo diferentes sujetos comparten los mismos conceptos–.)

[47] Obsérvese que los argumentos a favor de que invariantismo implica nativismo, y de que empirismo implica contextualismo, comparten una misma tesis nuclear –punto (4) de ambos argumentos–, a saber, que sin experiencias comunes (o, en el extremo, sin las mismas experiencias) sobre una cierta categoría, no es posible aprender un mismo e idéntico concepto. Y, aunque podría intentar rechazarse esta tesis, en ese caso la carga de la prueba caería del lado de quien lo hiciera, dado que tendría que mostrar cómo el mismo concepto puede adquirirse a partir de experiencias distintas y no conjuntamente equivalentes.

[48] En la medida en que este argumento puede extrapolarse a la comparación entre diferentes momentos de tiempo en la vida de un sujeto (en los que su biografía y experiencias sean distintas), este punto se puede generalizar afirmando que los conceptos no son invariantes ni entre individuos ni entre tiempos.

[49] Debo insistir en que el esquema siguiente (esto es, puntos [1] a [4]) no debe ser entendido como un argumento en favor del empirismo, sino como una explicación de uno de los motivos que conducen al empirista a asumir la premisa (1) del ARGUMENTO EiC –a saber, la tesis de que *los conceptos son adquiridos a partir de la experiencia*–.

[4] Luego, por inferencia hacia la mejor explicación, los conceptos son aprendidos.

Obviamente, este esquema de motivación podría ser puesto en cuestión. Por ejemplo, podría se replicar que aunque el mundo y las categorías presentes en él hubiesen cambiado –punto [2]–, es perfectamente posible que nuestros conceptos sobre tales categorías no lo hayan hecho, y sigan siendo los mismos. Esto es lo que parece que sucede en el caso de los *peligros* del mundo, que han cambiado aún cuando nuestras *fobias* y *temores* no lo hayan hecho. No obstante, ésta u otras críticas al esquema de argumentación [1]-[4] no resultan algo crítico para el ARGUMENTO EiC, en la medida en que el propósito de este esquema es únicamente explicar cuál podría ser el origen de la motivación del empirista quien, sin embargo, también podría simplemente asumir la premisa (1).

1.5.3 Invariantismo y empirismo son difícilmente compatibles

En las dos secciones anteriores he defendido la existencia de las dos relaciones de implicación siguientes entre los extremos de los debates nativismo *vs* empirismo e invariantismo *vs* contextualismo:

$$\text{Invariantismo} \rightarrow \text{Nativismo}$$

$$\text{Empirismo} \rightarrow \text{Contextualismo}$$

No obstante, cuando se examinan en detalle los argumentos allí empleados se observa que ambos hacen uso en sus puntos (4) de las dos relaciones de equivalencia asumidas al comienzo de esta sección 1.5. De hecho, la estructura "completa" de tales argumentos (incluyendo ese paso intermedio) sería la siguiente:

$$(\text{Invariantismo} \rightarrow \neg\,\text{Empirismo}) \wedge (\neg\,\text{Empirismo} \leftrightarrow \text{Nativismo})$$

$$(\text{Empirismo} \rightarrow \neg\,\text{Invariantismo}) \wedge (\neg\,\text{Invariantismo} \leftrightarrow \text{Contextualismo})$$

Ahora bien, si consideramos por separado el primer coyunto de cada una de las expresiones anteriores apreciamos que en realidad son equivalentes[51]:

$$(\text{Invariantismo} \rightarrow \neg\,\text{Empirismo}) \equiv (\text{Empirismo} \rightarrow \neg\,\text{Invariantismo})$$

Aún más, ambas expresiones son a su vez equivalentes a una tercera, a saber:

$$(\text{Invariantismo} \rightarrow \neg\,\text{Empirismo}) \equiv \neg\,(\text{Invariantismo} \wedge \text{Empirismo})$$

Conforme a esta última expresión, invariantismo y empirismo (en sus versiones extremas[52]) serían aproximaciones difícilmente compatibles, en la medida en que no parece posible adoptar de manera consistente una postura que las combine a ambas[53].

[50] O, dicho de otro modo, ¿cómo es posible que sean innatos conceptos de categorías que solo han existido en un pasado reciente? (Una cuestión que, conforme avanzamos hacia atrás en el tiempo, se extiende a un número creciente de categorías –potencialmente, a todas ellas–.)

[51] Ésta es la razón por la cual decía que los puntos (4) de ambos argumentos comparten la misma tesis nuclear (*vid.* nota al pie 47 en este capítulo), a saber, porque una y otra son en realidad proposiciones equivalentes.

1.5.4 Postura nativista-contextualista

Llegados a este punto queda únicamente una última cuarta postura por examinar, a saber, la combinación de nativismo y contextualismo. Anteriormente (en las secciones 1.5.1 y 1.5.2) se ha sugerido cuáles podrían ser las motivaciones empiristas e invariantistas –cada una de ellas por separado–. Veamos ahora cuáles serían las motivaciones del nativista y del contextualista, y en qué medida éstas les conducen a posiciones más o menos compatibles.

En primer lugar, puede decirse que el *nativista* fundamenta su postura en la incapacidad mostrada por el empirismo para proporcionar una explicación no-circular de cómo los conceptos se adquieren. Por su parte, la motivación del *contextualista* podría decirse que surge en torno a la evidencia empírica de que nadie haya sido nunca capaz de proporcionar una definición de nada[54] –o, alternativamente, de que nadie haya sido capaz de indicar en qué podría consistir "seguir una regla" (Kripke 1982)–.

Dicho lo anterior, las motivaciones de ambas posturas –*nativista* y *contextualista*– parecen compatibles, esto es, pudiera ocurrir que todos los conceptos fueran innatos y dependientes del contexto. No obstante, se trata de una combinación complicada, pues implica asumir que toda dependencia contextual es innata (en tanto en cuanto que dicha dependencia contextual no es más que una parte de los conceptos, los cuales en este caso son concebidos como innatos). El principal problema de esto es que con ello se estaría asumiendo que el sistema cognitivo del sujeto anticipa de modo innato todo posible contexto, en la medida en que pueda anticipar el modo en que los conceptos dependerían de tales contextos.

En consecuencia, aún cuando la combinación de nativismo y contextualismo es una posibilidad teóricamente viable, resulta una opción mucho menos natural que las combinaciones invariantismo-nativismo y empirismo-contextualismo antes discutidas.

1.5.5 Recapitulación

Sobre la base de lo indicado en los apartados anteriores, las tres combinaciones de posturas compatibles en los debates nativista-empirista e invariantista-contextualista, serían las que se muestran en la Tabla 1.1.

[52] Conviene realizar esta salvedad, en la medida en que un invariantismo meramente intrasubjetivo sí podría ser compatible con el empirismo, en la medida en que para dicho tipo de invariantismo no es preciso que sujetos distintos compartan el mismo concepto *C*. No obstante, tal compatibilidad lo sería a costa de una de las dos principales ventajas del invariantismo sobre el contextualismo –a saber, su capacidad para explicar de manera fácil la comunicación exitosa entre dos individuos–.

[53] Es por esto que la casilla de la Tabla 1.1 (con las alternativas teóricamente viables en los debates nativista-empirista e invariantista-contextualista) correspondiente a la combinación de empirismo e invariantismo se encuentre vacía.

[54] Sobre la base de esta evidencia el contextualista sostendría que eso es debido a que los conceptos dependen de (o cambian con) el contexto en que se aplican. Esto último –a saber, que en muchas ocasiones los conceptos dependen del contexto– estaría a caballo entre ser una consecuencia de que no existen definiciones, y ser una evidencia empírica *per se*.

	Invariantismo	Contextualismo
Empirismo		×
Nativismo	×	?

Tabla 1.1. Alternativas viables para la combinación de debates nativista-empirista (sobre el origen de los conceptos) *e invariantista-contextualista* (sobre su grado de dependencia contextual). Las dos combinaciones "naturales" han sido marcadas con una aspa (×); mientras que la tercera combinación compatible, aunque con mayor dificultad, ha sido marcada con un signo de interrogación (?).

1.6. Resumen

En este capítulo he repasado cuáles son las diferentes posturas que pueden adoptarse acerca de la estructura interna, naturaleza, origen y grado de dependencia contextual de los conceptos. El propósito de la presente sección es hacer explícitos los presupuestos que asumiré como punto de partida, lo que permitirá acotar el alcance de mi trabajo. En cualquier caso, a la hora de adoptar una u otra postura, partiré siempre del hecho de que lo único que observamos de los conceptos es aquello que nos permiten hacer, y eso es, en primera instancia categorizar (Harnad 2005), y en segundo término realizar inferencias.

En cuanto a su estructura interna, se ha visto que la distinción entre conceptos primitivos y complejos permite explicar, sobre la base del principio de composicionalidad, tanto la sistematicidad como la productividad del pensamiento. En este caso, asumo la postura mayoritaria en filosofía de la mente y ciencia cognitiva, según la cual la mayor parte de los conceptos son complejos, y se encuentran constituidos por otros más simples –en último término, por conceptos primitivos/atómicos–.

En lo relativo a cuál podría ser su naturaleza, se han presentado las tres principales posturas existentes al respecto, las cuales conciben los conceptos, bien como representaciones mentales, bien como habilidades, bien como entidades abstractas. En relación a este punto, no doy por sentado que puedan existir, ni representaciones mentales, ni habilidades mentales, sino que me limitaré a subscribir aquella postura que proporcione una mejor explicación de los fenómenos considerados[55]. Finalmente, en el capítulo 5 de esta tesis me inclinaré hacia la conveniencia de optar por una visión de los conceptos como habilidades mentales, en el sentido de herramientas cognitivas empleadas por nuestra mente en tareas de categorización.

Con respecto al origen y grado de dependencia contextual de los conceptos, se ha mostrado que las dos posturas extremas que cabe adoptar en sus respectivos debates (a saber, nativismo-empirismo para el caso del origen de los conceptos, e invariantismo-contextualismo en cuanto a su grado de dependencia contextual) no son independientes entre sí,

[55] El hecho de que las dos primeras aproximaciones (conceptos como representaciones y como habilidades) sean compatibles con todas las distintas posturas que cabe adoptar con respecto al origen y grado de dependencia contextual de los conceptos, y el hecho de que incluso la tercera aproximación (conceptos como entidades abstractas) pueda reinterpretarse –aunque, eso sí, renunciando a algunos de sus compromisos objetivistas– de un modo parcialmente compatible con la visión de los conceptos como habilidades mentales, es lo que no obliga a adoptar ninguna posición de antemano con respecto a ellas.

sino que presentan relaciones cruzadas relevantes. En primer lugar, con respecto al debate invariantista-contextualista, encuentro bastante implausible la idea de que pueda haber conceptos invariantes, por lo que en este caso la postura que asumiré será de tipo contextualista, y será en su marco en el que articule mi propuesta en esta tesis. En segundo lugar, creo que en la dialéctica entre empirismo y nativismo existe la posibilidad de asumir el empirismo. Por un lado, las evidencias existentes no deciden por sí mismas el debate, por lo que parece prudente no asumir que existan elementos innatos salvo que se disponga de buenos argumentos para sostenerlo. En este sentido, el mejor argumento del innatismo es de tipo negativo –esto es, un argumento en contra de la posibilidad de que los conceptos puedan aprenderse[56]–, por lo que mi primer paso será evaluar (en el capítulo 6 de esta tesis) dicho argumento sobre la base de una asunción empirista. Si, como intentaré mostrar en ese capítulo, el empirista es capaz de proporcionar una respuesta satisfactoria a la crítica nativista, entonces se encontrará en una mejor situación que los partidarios del innatismo.

[56] O, más en particular, en contra de la posibilidad de que pueda haber una explicación no-circular de cómo los conceptos primitivos se adquieren.

This page intentionally left blank

Capítulo 2: Teorías sobre la estructura de los conceptos

¿Pues de qué modo está cerrado el concepto de juego? ¿Qué es aún un juego y qué no lo es ya? ¿Puedes indicar el límite?. –
Ludwig Wittgenstein (1953, §68)

2.1. Introducción

En el anterior capítulo 1 indicaba que –en el ámbito de la filosofía de la mente y ciencia cognitiva– se suele aceptar la tesis de que los conceptos (*complejos*) son estructuras constituidas por otros conceptos más básicos. También suele asumirse, en virtud del principio de composicionalidad, que todo concepto complejo hereda su significado de sus conceptos constituyentes. La aproximación alternativa sería el *atomismo conceptual*, conforme al cual: (i) los conceptos carecerían de estructura interna, y (ii) su contenido estaría fijado por las relaciones causales que mantienen con las cosas del mundo (Fodor 1998; Margolis 1998; Millikan 2000). En todo caso, aún dentro de la corriente para la cual los conceptos tienen estructura interna, existe una considerable controversia con respecto a qué aproximación a esa estructura interna de los conceptos resulta más adecuada para caracterizarlos (Margolis y Laurence 2011a).

Primeramente está la *teoría clásica*, según la cual la mayoría de los conceptos pueden concebirse como definiciones que codifican las condiciones necesarias y conjuntamente suficientes para su aplicación. No obstante, a pesar de las ventajas de la teoría clásica (a saber, su simplicidad y potencia explicativa a la hora de dar cuenta de fenómenos tales como adquisición de conceptos, categorización, inferenciación, justificación epistémica, entre otros), la teoría presenta importantes objeciones en su contra como, por ejemplo, su incapacidad para proporcionar definiciones exitosas de casi nada (Wittgenstein 1953; Gettier 1963; Fodor 1981a), los problemas de la noción de analiticidad (Quine 1951) o la existencia de fenómenos de tipicalidad (Rosch y Mervis 1975), además de otros problemas más generales –y compartidos por algunas otras teorías sobre la estructura de los conceptos–, como el problema de la ignorancia y el error, o los debidos a la vaguedad conceptual.

Debido a todos estos problemas han ido surgiendo teorías alternativas a la clásica, con el propósito de explicar cuál es la naturaleza interna de los conceptos sin caer en los problemas mencionados. Por un lado está la *teoría de prototipos*, que concibe a los conceptos como representaciones complejas con estructura probabilística, en base a las que un cierto objeto cae dentro de un determinado concepto si satisface un suficiente número de propiedades asociadas a dicho concepto (Wittgenstein 1953; Rosch y Mervis 1975; Rosch 1978; Hampton 1979). En todo caso, la teoría de prototipos comparte con la teoría clásica la asunción de que los objetos se clasifican bajo una cierta categoría en virtud de su similaridad –esto es, atributos compartidos– con alguna especificación de esa categoría, razón por la cual en ocasiones se las describe –juntamente con la teoría de ejemplares– como *enfoques basados en similaridades* (Medin 1989; Komatsu 1992). Ahora bien, las aproximaciones basadas en similaridades se enfrentan a lo que Machery (2009, p. 85) llama el "problema de selección" –que no es otro que la séptima objeción de Goodman (1972)–, consistente en que, en ausencia de unos principios que determinen qué propiedades cuentan como relevantes y cuál es la importancia de cada una de ellas, los modelos basados en similaridades resultan inútiles[1]. Con respecto al resto de las objeciones a la teoría de prototipos, algunas ya estaban presentes en la teoría clásica –como el problema de la ignorancia y el error–, mientras que otras son específicas suyas, tales como la existencia de fenómenos de tipicalidad inesperados (Armstrong *et al.* 1983), la ausencia de juicios de tipicalidad para ciertos conceptos (Fodor 1981a), o la dificultad que presenta la teoría para explicar el fenómeno de la composicionalidad (Osherson y Smith 1981).

Una tercera aproximación dentro de los enfoques basados en similaridades es la conocida como *teoría de ejemplares*, según la cual un concepto no es más que un conjunto de ejemplares o, dicho de otro modo, el cuerpo de conocimiento sobre las propiedades de los miembros individuales de tal concepto (Medin y Schaffer 1978). En tal caso, la determinación de si algo cae, o no, bajo una cierta categoría tendrá lugar por medio del cálculo de su similaridad con respecto a todos los ejemplares previamente encontrados, y su posterior asignación a la categoría asociada al ejemplar más próximo. En cuanto a sus problemas, la teoría de ejemplares adolece del mismo problema de selección que la teoría de prototipos –ambas son aproximaciones basadas en similaridades–. Por otro lado, entre las críticas específicas de esta teoría destacan la necesidad de explicar la presencia de información sobre las tendencias centrales de las categorías –en el caso de que los conceptos fuesen meros conjuntos de ejemplares– (Komatsu 1992). o la dificultad de aceptar (desde un punto de vista computacional) que para clasificar un objeto haya que computar su similaridad con todos los ejemplares previamente observados.

Otra aproximación alternativa es la conocida como *teoría-teoría*, la cual caracteriza los conceptos en términos de las relaciones que esos conceptos mantienen entre sí (Carey 1985, 2009; Murphy y Medin 1985; Keil 1989; Gopnik y Meltzoff 1997), de un modo similar a cómo los términos de una teoría científica se relacionan entre sí. De nuevo, esta concepción presenta problemas, algunos compartidos con otras concepciones –como el de la ignorancia y el error–, y otros específicos suyos como, por ejemplo, la pobre comprensión que tenemos de cómo ocurre la emergencia de nuevas teorías (o la transición de una teoría a otra).

[1] La cuestión de por qué los conceptos tan solo representan algunas de las numerosas propiedades presentes en los miembros de sus categorías es algo en lo que, desde el ámbito de la psicología, se ha insistido de manera recurrente (Goodman 1972; Smith y Medin 1981; Medin 1989).

Finalmente, el hecho de que ninguna de las teorías anteriores haya sido capaz de formular un modelo que proporcione una explicación a la formación y aplicación de todo concepto ha conducido a muchos autores a aceptar que ninguna de esas teorías por separado podrá nunca proporcionar una explicación exitosa de todos estos fenómenos. Este "nuevo consenso" –así referido por Bloch-Mullins (2017)– con respecto al hecho de que las diferentes teorías anteriores (prototipos, ejemplares y teorías) no tienen por qué ser incompatibles entre sí, ha cristalizado en tres aproximaciones principales, a saber, pluralismo, hibridismo y eliminativismo. Las dos primeras sostienen que los conceptos tienen múltiples estructuras asociadas, bien como clases de conceptos distintas operando de manera específica en cada tarea cognitiva –*pluralismo*– (Piccinini y Scott 2006; Weiskopf 2009a), bien como diferentes partes de un mismo concepto que operarían de modo simultáneo en toda tarea cognitiva –*hibridismo*– (Smith *et al.* 1974; Osherson y Smith 1981; Nosofsky *et al.* 1994; Anderson y Betz 2001). Por su parte, el *eliminativismo* (Machery 2009) acepta la tesis pluralista de que las categorías tienen asociadas diferentes clases de conceptos con pocas propiedades en común, que cumplirían distintas funciones cognitivas, pero –a diferencia del pluralismo– concluye de ello que la noción de concepto es inútil, por lo que la ciencia cognitiva debería eliminarla de su vocabulario teórico y, ya sin ella, dedicarse al estudio de los prototipos, ejemplares y teorías.

En los apartados siguientes presentaré los principios, motivación y principales puntos fuertes asociados a cada una de estas teorías, tras lo cual repasaré las más importantes críticas y objeciones recibidas por cada una.

2.2. Enfoques basados en definiciones

El primer gran grupo de teorías sobre la estructura de los conceptos podría describirse como perspectiva definicional, o *teorías basadas en definiciones*[2], que incluiría tanto a la *teoría clásica*, como a aquellas variaciones de la teoría clásica original surgidas con objeto de dar respuesta a algunos de sus principales problemas –y que, siguiendo a Laurence y Margolis (1999), aquí serán referidas como *teorías neo-clásicas*–.

2.2.1 Teoría clásica

La teoría clásica se remonta a la antigua filosofía griega, encontrándola en los diálogos platónicos *Eutifrón*, *Lisis* y *Laques* –cuando Sócrates investiga la naturaleza de conceptos tales como la piedad, la amistad y el valor–, y mantuvo su hegemonía en filosofía y psicología[3] hasta la segunda mitad del siglo XX. Y, aunque en la actualidad haya perdido buena parte de su dominancia histórica, la teoría clásica sigue presentando un gran interés en la medida en que todas las otras teorías sobre la estructura de los conceptos son –de una u otra forma– reacciones o extensiones a la teoría clásica.

[2] Los enfoques basados en definiciones a veces son también referidos como *teorías basadas en reglas* (Smith y Sloman 1994; Ashby *et al.* 1998).

[3] Véanse, por ejemplo, los trabajos de Hull (1920) y Bruner *et al.* (1956) –en psicología– y de Katz y Fodor (1963) –en filosofía–.

Para la teoría clásica la mayoría de los conceptos tienen una estructura definicional, razón por la cual esta aproximación también recibe el nombre de *perspectiva definicional*. La idea es que un concepto codifica las condiciones necesarias y conjuntamente suficientes para su aplicación que, para el caso de los conceptos complejos, estarían determinadas por los conceptos más simples (o *rasgos*) de que tales conceptos complejos se componen. Bajo este enfoque, el concepto SOLTERO podría decirse compuesto de los conceptos VARÓN, ADULTO, y NO-CASADO, en la medida en que cualquier cosa que satisfaga esas condiciones sea un soltero.

La ya mencionada hegemonía que ha mantenido históricamente la teoría clásica ha sido debida a su gran simplicidad y potencia explicativa con respecto a un gran número de cuestiones clave:

— *Adquisición de conceptos*: si los conceptos complejos no fuesen más que constructos compuestos de otros más simples (que son las condiciones necesarias y suficientes para su aplicación), entonces la adquisición de un concepto se reduciría al ensamblaje de los rasgos que constituyen su definición. Esto conduce, en último término, a la posibilidad de sostener que todo concepto complejo puede definirse en base a un repertorio relativamente pequeño de conceptos sensoriales, tal y como ha sido defendido tradicionalmente desde el empirismo (Locke 1690; Carnap 1932).

— *Categorización*: la categorización de algo bajo un cierto concepto se reduciría a la comprobación de que ese algo satisface los rasgos que definen ese concepto. Y, si los constituyentes últimos de los conceptos expresasen propiedades sensoriales, esa verificación debería resultar poco problemática.

— *Inferenciación analítica*: puesto que toda inferencia analítica está basada en el significado de sus elementos constituyentes, de atribuir un concepto *C* a un objeto *o* se podrá inferir la atribución –al objeto *o*– de todos los rasgos que constituyen la definición de *C*.

— *Justificación epistémica*: un sujeto estará justificado epistémicamente para pensar que algo cae bajo una cierta categoría tras comprobar que los rasgos que definen ese concepto son satisfechos por el objeto considerado.

— *Determinación de referencias*: trivialmente, los conceptos refieren a aquellas cosas del mundo que satisfacen sus definiciones.

Ahora bien, aún a pesar de los puntos fuertes anteriores, existen importantes dudas con respecto a si la teoría clásica constituye una aproximación adecuada para caracterizar de manera exitosa la noción de concepto. La razón es que existen significativas críticas y objeciones a esta teoría, entre las que destacan las siguientes[4]:

[4] En ocasiones también se plantea en contra de la teoría clásica la crítica de que la estructura definicional que asume no se manifiesta en una variedad de contextos experimentales en donde sí que cabría esperarla –*problema de la realidad psicológica*–. O, dicho de otro modo, la complejidad que manifiestan los conceptos en psicología experimental (en términos de carga de procesamiento) no parece depender de la complejidad de sus definiciones asociadas (Kintsch 1974; Fodor *et al.* 1980). No obstante, ésta es una objeción menor, en la medida en que los resultados de esos experimentos pueden explicarse sin tener que recurrir al abandono de la perspectiva definicional (Laurence y Margolis 1999, p. 18).

(1) *Contra la existencia de definiciones*: la principal crítica contra la teoría definicional es que no tenemos una definición para casi ningún concepto –sobre todo si, conforme a los principios empiristas, tales definiciones deben estar fundadas en elementos perceptuales–. En ocasiones esta objeción es referida como el "problema de Platón" (Laurence y Margolis 1999), pues éste en sus diálogos ya puso de manifiesto la dificultad para encontrar definiciones de casi cualquier cosa. Desde entonces, y aún a pesar del esfuerzo de muchos filósofos por encontrar definiciones para conceptos tales como CONOCIMIENTO, BONDAD, VERDAD, etc., ninguna de ellas resulta convincente e incontrovertible[5]. Aún peor, el problema para encontrar definiciones se extiende más allá de conceptos "filosóficos" como los anteriores, alcanzado también a los conceptos "comunes", conforme mostraron Wittgenstein (1953) y Fodor (1981a) con sus discusiones relativas a la posibilidad de definir, o no, los conceptos JUEGO y PINTAR, respectivamente[6]. Y, aunque podría pensarse que la falta de definiciones admisibles se debe a que la tarea resulta mucho más difícil de lo que se había supuesto, la situación se parece más a la retratada por Platón en sus diálogos, en el sentido de que las definiciones propuestas nunca parecen estar exentas de contraejemplos, lo que conduce a la sospecha de que nuestros conceptos carecen de estructura definicional.

(2) *Contra la noción de analiticidad*: en este caso la cuestión es que, sin ejemplos de definiciones concretas ni evidencias psicológicas que ofrecer, el principal sustento de la teoría clásica es su capacidad para explicar la inferenciación analítica. No obstante, la crítica de Quine (1951) a la noción de analiticidad, y su conclusión de que la confirmación de las afirmaciones tiene lugar de manera holista[7] y, por consiguiente, no hay condiciones de confirmación que puedan establecerse a priori, constituye una seria dificultad para el segundo apoyo de la teoría clásica. En este caso el problema es doble, puesto que acabar con la distinción analítico-sintético supone, no solo acabar con la noción de inferenciación *analítica*, sino también con el uso de la noción de analiticidad –en el sentido neo-positivista– en el ámbito de la justificación epistémica.

(3) *Existencia de fenómenos de tipicalidad*: la idea de parecido de familia introducida por Wittgenstein (1953) en su discusión de si es posible definir o no el concepto JUEGO arrojaba dudas –desde una perspectiva teórica– acerca de la posibilidad de que los conceptos fuesen definiciones. Dos décadas después, en la década de 1970, dichas dudas fueron confirmadas cuando desde el ámbito de la psicología empírica se identificó la existencia de efectos de tipicalidad en un gran número de conceptos, entre los que cabe destacar los siguientes:

[5] Incluso, la prometedora definición de CONOCIMIENTO como CREENCIA VERDADERA y JUSTIFICADA presenta significantes problemas, tal y como mostró Gettier (1963) con sus conocidos problemas.

[6] Incluso la definición de un concepto en apariencia tan claro como SOLTERO –a saber, como VARÓN, ADULTO, y NO-CASADO– no se encuentra libre de contraejemplos (Fillmore 1982; Lakoff 1987).

[7] En contra de la asunción neopositivista de que las relaciones analíticas son tautologías originadas en las convenciones del lenguaje (Carnap 1947), lo cual permitiría su conocimiento a priori mediante el mero análisis lingüístico de las condiciones que verifican cada afirmación.

— Las personas ordenan sin dificultad –y de modo consistente– un conjunto de objetos en términos de cuán buenos son como ejemplos de una cierta categoría (Rosch 1973). Por ejemplo, la clasificación de tipicalidad de los miembros de la categoría FRUTA produce una ordenación muy concreta (*manzana - ciruela - piña - fresa - higo - aceituna*), que no depende tan siquiera de la familiaridad de los sujetos con los objetos considerados.

— Las personas juzgan que un miembro de una categoría es tanto más típico o representativo cuantos más rasgos tiene en común con otros miembros de esa categoría y menos rasgos comparte con los miembros de otras categorías[8] (Rosch y Mervis 1975). Por ejemplo, en la categoría PÁJARO, la tipicalidad de *gorriones*, *águilas* y *pollos* es distinta y decreciente.

— Las personas categorizan más rápido y con menos errores los objetos más típicos de una categoría, que aquellos otros menos típicos (Rosch 1973; Smith et al. 1974).

El problema es que la teoría clásica, con su definición de concepto en términos de condiciones necesarias y suficientes, era incapaz de explicar la presencia de este tipo de fenómenos[9]. O, dicho de otro modo, si los conceptos fuesen definiciones entonces todo objeto que cayese bajo una cierta definición debería ser un ejemplo igual de "bueno" que cualquier otro miembro de esa categoría, justo en contra de lo que sugieren los fenómenos de tipicalidad (a saber, que no todos los miembros de una categoría son igualmente representativos de esa categoría).

Finalmente, la teoría clásica comparte algunas de las dificultades que encontramos también presentes en muchas de las otras teorías sobre la estructura de los conceptos que se han propuesto como alternativa a la teoría clásica, y entre los que destacan el problema de la ignorancia y el error, así como el debido a la vaguedad conceptual. Con respecto al primero –*problema de la ignorancia y el error*–, los argumentos planteados por Putnam (1970) y Kripke (1980) en contra de la teoría descriptivista de la referencia socavaban – cuando se aplicaban al caso de nombres propios y términos de género natural– la teoría clásica sobre la estructura de los conceptos, en la medida en que ésta no era más que descriptivismo aplicado a los conceptos. Por un lado, es posible tener un concepto y, al mismo tiempo, estar equivocados sobre las propiedades atribuidas a sus instancias, en cuyo caso esas propiedades no podrían ser parte de la definición de ese concepto. En esta situación aceptamos que, aun pudiendo estar equivocados con respecto a dichas propiedades (y, por ello, no disponiendo de la definición correcta de ese concepto), tenemos el concepto en cuestión –*argumento desde el error*–. Por otro lado, podemos ignorar muchas de las propiedades de un determinado concepto (tal y como ha ocurrido en el pasado, y seguramente sigue ocurriendo en el presente, para muchas categorías), y aún así estar dispuestos a aceptar que tenemos dicho concepto –*argumento desde la ignorancia*–. Estos

[8] Como ocurre, por ejemplo, con la categoría PÁJARO, en donde la tipicalidad de *gorriones*, *águilas* y *pollos* es distinta y decreciente.

[9] Tal incapacidad dio lugar –como reacción– a la primera formulación de la teoría de prototipos (Rosch y Mervis 1975; Rosch 1978).

dos argumentos (desde el error y la ignorancia) constituyen razones de peso para sostener que podemos tener un concepto aún sin disponer de las condiciones necesarias y suficientes para su aplicación –es decir, su definición–, lo que explicaría la posibilidad de realizar categorizaciones erróneas en base al mismo (esto es, clasificar bajo él cosas que no son miembros de esa categoría, y no incluir otras cosas que sí lo son).

En cuanto al segundo –*problema de la vaguedad conceptual*–, la teoría clásica considera que los conceptos tienen extensiones determinadas, pues la definición de un concepto conduciría a unas categorizaciones definidas. Sin embargo, muchos de nuestros conceptos son borrosos o inexactos (en el sentido de que contienen siempre una cierta cantidad de indeterminación). Considérese, por ejemplo, el caso del concepto MOBILIARIO y el dilema de decidir si las *alfombras* pertenecen o no a él (Medin 1989); o el caso de los *tomates* y el debate de si deben clasificarse bajo la categoría FRUTA (Smith y Medin 1981). El hecho de que la teoría clásica busque rasgos definidores no ambiguos constituye un problema a la hora de explicar este tipo de fenómenos[10].

2.2.2 Teorías neo-clásicas

Muchas de las anteriores objeciones a la teoría clásica han intentado responderse desde el propio marco definicional, lo que ha dado lugar a propuestas que podrían calificarse como *teorías neoclásicas*. Por ejemplo Jackendoff (1983), sin renunciar a la existencia de unas condiciones necesarias, estaría dispuesto a aceptar (i) condiciones graduadas y (ii) condiciones estereotípicas con excepciones –que permitirían explicar tanto los fenómenos de tipicalidad, como las objeciones de Wittgenstein en contra de la posibilidad de proporcionar definiciones–. Por su parte Pinker (1989) considera que las definiciones podrían consistir en estructuras híbridas que combinasen elementos universales y recurrentes –condiciones necesarias–, con conocimiento sobre el mundo real. Nosofsky y colaboradores (1994) plantean una propuesta similar, basada en la combinación de reglas –condiciones necesarias– y excepciones a esas reglas que operarían sobre la base de ejemplares específicos almacenados en la memoria; y Sloman (1996) argumenta a favor de la existencia de dos sistemas de razonamiento, a saber, uno basado en reglas y otro de tipo asociativo –este último en base a relaciones de similaridad y contigüidad temporal–. Finalmente, también cabe destacar un cierto renacimiento bayesiano de los enfoques basados en reglas, de la mano de Tenenbaum y Griffiths (2001), cuando unifican en un mismo marco bayesiano (a) una caracterización del modelo de similaridad de Tversky (b) por medio de la teoría de conjuntos.

Ahora bien, dejando de lado las diferencias entre cada variante en particular, puede decirse que la teoría neo-clásica se caracteriza por sostener que los conceptos tienen *definiciones parciales* que codifican el conjunto de condiciones necesarias que algo deberá

[10] En ocasiones la teoría clásica responde a esta dificultad introduciendo en las condiciones de aplicación de los conceptos ese carácter borroso. Ahora bien, con ello queda abierta a la primera de las críticas referidas, a saber, la no-existencia de definiciones, en la medida en que unas condiciones de aplicación no-definidas no constituyen una definición.

satisfacer para caer bajo su extensión pero que, no obstante, dichas condiciones no son suficientes para que ocurra la clasificación bajo ese concepto[11].

En todo caso, la motivación de quienes se adscriben a los enfoques neo-clásicos no suele ser tanto la de rescatar o preservar la teoría clásica, como sí la de recurrir a las definiciones parciales por su capacidad para explicar ciertos fenómenos lingüísticos presentes en construcciones causativas[12], polisemias, alternancias sintácticas y adquisición léxica (Laurence y Margolis 1999, p. 53).

Para terminar, las teorías neo-clásicas no están libres de problemas, algunos de ellos compartidos con la teoría clásica (como, por ejemplo, el *problema de la ignorancia y el error*), y otros específicos de ella. En cuanto a estos últimos, posiblemente el más relevante sea el *problema de incompletitud de las definiciones parciales*, las cuales deben ser rellenadas con algún tipo de "completador" cuando esas definiciones parciales son aplicadas en tareas de categorización, determinación de la referencia, etc. Sin embargo, al considerar cómo puede tener lugar la compleción de la definición parcial, ninguna de las alternativas está libre de problemas dado que: (i) si la compleción da lugar a la conversión de la definición parcial en una definición completa, caeríamos de nuevo en todos los problemas propios de la teoría clásica; pero, (ii) si tras su compleción la definición sigue manteniendo un carácter parcial, entonces no está claro cómo una definición parcial podría producir una categorización o fijación de la referencia concretas. O, dicho de otro modo, para explicar cómo tienen lugar fenómenos tales como los de categorización o determinación de la referencia las definiciones parciales de la teoría neo-clásica han de convertirse en definiciones completas, lo que supone una vuelta a la teoría clásica y, por consiguiente, a todos sus problemas.

2.3. Enfoques basados en similaridades

Frente a los enfoques basados en definiciones (o reglas), el segundo gran tipo de teorías sobre la estructura de los conceptos se encuentra basado en la noción de similaridad, e incluiría tanto a la *teoría de prototipos* como a la *teoría de ejemplares*[13]. Las aproximaciones basadas en similaridades se caracterizan por sostener que la clasificación de un objeto bajo una cierta categoría tiene lugar en virtud de los rasgos que ese objeto comparte con el

[11] Obsérvese que la mera aceptación por parte de ciertos tipos de pluralismo de la noción de definición como una de las múltiples estructuras de conceptos que pueden operar en los distintos dominios (Pinker y Prince 1996; Ashby *et al.* 1998) no las convierte *per se* en aproximaciones neo-clásicas.

[12] Por ejemplo, Jackendoff (1989) considera que construcciones causativas tales como *x persuadió a y de que P* dan lugar a inferencias del tipo *y llegó a creer que P* porque uno de los conceptos presentes en la primera (PERSUADIR, en este caso) tiene una estructura en la que PROVOCAR UNA CREENCIA es una definición parcial suya. Dicha definición parcial constituiría una condición necesaria –que no suficiente– para la aplicación del concepto PERSUADIR, razón por la cual la inferencia anterior es posible.

[13] Aunque algunos autores también incluyen a la teoría clásica dentro de los enfoques basados en similaridades (Medin 1989; Komatsu 1992), en este trabajo se ha optado por la práctica habitual de incluir en estos enfoques únicamente a las teorías de prototipos y ejemplares (Machery 2009; Bloch-Mullins 2017).

concepto asociado a dicha categoría. (Cuanto mayor sea el número de atributos compartidos, tanto más fácil será que el objeto sea clasificado bajo la categoría.)

La principal motivación de los enfoques basados en similaridades es caracterizar la noción de parecido de familia (Wittgenstein 1953) mediante la formulación de modelos que la articulen y, con ello, permitir explicar los fenómenos de tipicalidad identificados por Rosch y Mervis (1975), lo cual no era posible desde el ámbito de las teorías basadas en definiciones.

2.3.1 Teoría de prototipos

La teoría de prototipos fue la primera propuesta que surgió con objeto de conciliar los resultados empíricos que demostraban la existencia de efectos de tipicalidad, y su aparición supuso el fin de la hegemonía de la teoría clásica[14]. No obstante, y puesto que no existe una sola versión de esta teoría, el propósito de la presente sección será el de presentar el principal núcleo común de sus diferentes versiones.

La teoría de prototipos –también llamada *perspectiva probabilista* (Medin 1989) o *visión de parecidos de familia* (Komatsu 1992)–, sobre la base de la evidencia empírica que muestra que las categorías son borrosas o imprecisas, sostiene que éstas pueden organizarse en torno a conjuntos de atributos correlacionados. Esos rasgos –relacionados de manera estadística– conformarían una representación ideal (por ejemplo, por medio de una tendencia central[15]) que resume las propiedades características de ese concepto, y a la cual se da el nombre de *prototipo*. Por tanto, para la teoría de prototipos los conceptos son estructuras estadísticas que codifican los atributos que sus miembros acostumbran a tener[16].

Dado el carácter probabilístico de la teoría será posible –e incluso habitual– que algunos miembros de una categoría no presenten –o instancien– una o varias de las propiedades representadas por el concepto asociado a esa categoría. Con ello, los rasgos de los conceptos dejan de tener el carácter necesario que tenían en la teoría clásica, y mientras en que aquélla la categorización de algo exigía la satisfacción de todas las condiciones suficientes, en el caso de la teoría de prototipos bastará con exigir la satisfacción de un número suficiente de esos atributos (Rosch 1978; Hampton 1979). Se trata, por lo tanto, de un enfoque en el que la pertenencia de los objetos a las categorías no se encuentra de-

[14] En todo caso, y aún cuando la propuesta de concebir los conceptos por medio de prototipos –como alternativa a la teoría clásica, con objeto de explicar los fenómenos de tipicalidad– nace del trabajo de Rosch (1975); la idea de *prototipo*, en el sentido de un esquema que es abstraído de un conjunto de ejemplos, es previa, tal y como muestran los trabajos de Attneave (1957) y, sobre todo, los de Posner y colaboradores (Posner y Keele 1968; Posner 1969).

[15] No obstante, la *media* –o *tendencia central*– no es el único modo en que puede caracterizarse la noción de prototipo. De hecho, otras alternativas serían el uso de: (a) una *caricatura idealizada* –o *estereotipo*– que distinga máximamente la categoría considera del resto de categorías; (b) la *moda*, esto es, el ejemplar –o combinación de ejemplares– más frecuente; o (c) los *rasgos modales* (es decir, rasgos más frecuentes) en el conjunto de instancias consideradas (Kruschke 2005, p. 187).

[16] La idea de fondo es que las personas –en base a su propia experiencia– abstraen las tendencias centrales de los ejemplares observados de una cierta categoría, a partir de las cuales se formarían los prototipos que aglomeran toda esa información.

terminada –en el sentido de bien definida–, lo que ha llevado a hablar tanto de la *indeterminación de la pertenencia a categorías* (Mervis y Rosch 1981), como de una *estructura graduada de los conceptos* (Barsalou 1983; 1987), con objeto de referir tanto al fenómeno de que esa pertenencia no es una cuestión de todo o nada (sino graduada), como al hecho de que la contribución de cada atributo a esa pertenencia también es algo graduado.

Todo ello permitía a la teoría de prototipos, no solo explicar gran parte de las evidencias empíricas relativas a los efectos de tipicalidad[17] sino, además, dar cuenta de muchas más inferencias que las que permitía la teoría clásica pues, mientras que la teoría clásica únicamente permitía inferencias demostrativas, con la teoría de prototipos se abre la puerta a la posibilidad de inferencia fiables pero falibles. Por otro lado, relaciones de pertenencia graduadas también explicarían por qué las fronteras de muchos conceptos parecen ser borrosas (Kamp y Partee 1995), de manera que la teoría de prototipos estaría libre del problema de la vaguedad conceptual del que adolecían las propuestas clásicas basadas en definiciones.

En cuanto a cómo ocurrirían los fenómenos de categorización, las representaciones ideales con que se identifican los prototipos determinarán la clasificación de un cierto objeto bajo una u otra categoría. Esto tendría lugar mediante un proceso de comparación entre ese objeto y los prototipos de los conceptos tentativos considerados, tras lo cual dicho objeto se clasificaría bajo la categoría a cuyo prototipo fuera más similar (Hampton 1998, 2006). Esto supone el paso de las condiciones necesarias y suficientes de la teoría clásica, a una concepción holista del significado, en donde la función de clasificación –aún siendo determinista (como también lo era en el caso de las definiciones)– no solo dependerá del prototipo de la propia categoría, sino también de los prototipos de cualesquiera otras categorías relevantes.

No obstante, y aún a pesar de sus ya mencionadas ventajas frente a los enfoques basados en definiciones, la teoría de prototipos tampoco está libre de objeciones. Por un lado, comparte con las teorías clásicas el *problema de la ignorancia y el error* que, aunque puede ser respondido diciendo que si algo no encaja con el prototipo de una categoría entonces es que no cae bajo dicho concepto, el precio a pagar por ello es no dejar espacio a la posibilidad de que un concepto sea mal aplicado (Laurence y Margolis 1999, p. 34). Por otro lado, en cuanto a las objeciones planteadas específicamente en contra de la teoría de prototipos cabe destacar las siguientes:

— *Existen conceptos bien definidos que presentan efectos de tipicalidad*: conforme hemos indicado, la teoría de prototipos concibe a los conceptos como estructuras estadísticas que explicarían los efectos de tipicalidad que se han identificado empíricamente en muchos conceptos. Por ello, se esperarían fenómenos de tipicalidad en toda aquella categoría cuyo concepto estuviera codificado en términos de relaciones probabilísticas, pero no en las categorías cuyos conceptos estuviesen

[17] Como, por ejemplo, los siguientes: (a) existencia de juicios graduales sobre la tipicalidad de los miembros de una cierta categoría; (b) correlación entre la tipicalidad atribuida a un ejemplar, y la frecuencia de los rasgos presentes en él –y asociados a esa categoría–; (c) distintos tiempos de respuesta en tareas de categorización en función de la frecuencia de los rasgos presentes en los ejemplares considerados; (d) correlación inversa entre errores de clasificación y tipicalidad del objeto clasificado.

bien definidos. El problema es que eso no es así, pues tales efectos de tipicalidad también han sido identificados en conceptos con buenas definiciones (Bourne 1982; Armstrong *et al.* 1983) como, por ejemplo, ABUELA o NÚMERO-PAR[18]. Esto constituye un serio punto débil en la principal evidencia a favor de la teoría de prototipos, pues abre la puerta a la posibilidad de que los fenómenos de tipicalidad se deban a cómo los sistemas de categorización operan, y no a la estructura de los conceptos (Osherson y Smith 1981).

— *Existen conceptos que no tienen asociados juicios de tipicalidad*: esta objeción es la inversa de la anterior, pues supone no encontrar fenómenos de tipicalidad allí en donde sí son esperados. El problema tiene su origen en el hecho de que muchos conceptos no tienen asociados juicios de tipicalidad, pues las personas no son capaces de determinar cuáles son sus tendencias centrales. Por ejemplo, aún cuando puede haber un prototipo de CIUDAD –como Roma o Londres– e incluso de CIUDAD-AMERICANA –como Nueva York o Los Ángeles–, en cambio seguramente no exista un prototipo de CIUDAD-AMERICANA-EN-LA-COSTA-ESTE-AL-SUR-DE-TENNESSEE[19] (Fodor 1981a). Además, también parece posible tener un concepto aún cuando no se conozca ningún rasgo suyo estadísticamente significativo.

— *Existen efectos de tipicalidad en conceptos derivados de objetivos asociados, no con el prototipo de tales categorías, sino con ideales que maximizan el logro de esos objetivos*: en este caso el problema es debido a la identificación de efectos de tipicalidad en las categorías derivadas de objetivos (propuestas por Barsalou (1983) para explicar cómo las personas construyen categorías instrumentales, por ejemplo COSAS-PARA-VENDER-EN-UN-GARAJE) para el logro de ciertos objetivos. En principio, si tales conceptos tuvieran un prototipo[20], la tipicalidad de sus miembros debería determinarse con respecto a dicho prototipo. Sin embargo, los estudios de Barsalou (1983; 1985; 1987) mostraban que las categorías derivadas de objetivos presentaban efectos de tipicalidad basados en similaridades relativas, no a una tendencia central o prototipo, sino al ideal que mejor permite alcanzar esos objetivos. Así, por ejemplo, para el caso del concepto COSAS-QUE-COMER-EN-UNA-DIETA la proximidad (esto es, similaridad) entre los ejemplos considerados y el ideal de cero calorías era lo que determinaba las puntuaciones de tipicalidad.

— *Dificultad para explicar la composicionalidad de los conceptos*: una de las principales objeciones en contra de la teoría de prototipos es su aparente incapacidad para ex-

[18] En este último caso, aún cuando los sujetos objeto de estudio respondían negativamente a la cuestión de si la categoría NÚMERO-PAR tenía grados, en cambio sí tendían a ordenar sus miembros diciendo que algunos eran mejores ejemplos de la categoría que otros (por ejemplo, que el número 8 era un mejor número par que el número 34), e identificaban –como números pares– con mayor rapidez a aquéllos a los que habían atribuido una mayor tipicalidad.

[19] Entre los conceptos que carecen de prototipo se encontrarían todos los conceptos sin instancias (como, por ejemplo, FILÓSOFOS-DEL-SIGLO-XXII), y también los conceptos con extensiones demasiado heterogéneas (como LIBRO-O-PUERTA o NO-MESA).

[20] Y no cayeran en el problema descrito en el punto anterior (esto es, en el problema de no disponer un prototipo por tratarse de conceptos con extensiones demasiado heterogéneas).

plicar cómo se pueden componer conceptos, razón por la cual suele llamárselo *problema de composicionalidad* de los prototipos. Según indicábamos en el capítulo anterior, el principio de composicionalidad resultaba clave para explicar tanto la productividad como la sistematicidad del pensamiento. Las primeras propuestas sugerían que las extensiones graduadas podían caracterizarse mediante conjuntos borrosos (Rosch y Mervis 1975; Oden 1977), pero pronto surgieron casos que mostraban que los conceptos así compuestos no siempre equivalían a lo que cabría esperar (Osherson y Smith 1981). Por ejemplo, el concepto conjuntivo MANZANA-RAYADA no equivale a la intersección de las extensiones de sus dos conceptos asociados[21]. Otra crítica relacionada es la de que no siempre el prototipo de un concepto complejo es una función de los prototipos de sus conceptos constituyentes. Ése es el caso del concepto PET FISH (en inglés), cuyos rasgos –los de los peces dorados– no resultan de la combinación de los rasgos de PET y FISH, en la medida en que el prototipo de PET FISH tiene poco que ver con los prototipos de sus partes constituyentes (Fodor 1998, pp. 102-108). Y, a pesar de que se han propuesto múltiples respuestas que intentan compatibilizar el principio de composicionalidad con la teoría de prototipos (Smith y Osherson 1984; Hampton 1987, 1991, 1997a; Smith *et al.* 1988; Kamp y Partee 1995; Costello y Keane 2000), eso no ha servido para cerrar un debate que a día de hoy continúa aún abierto entre aquellos que esgrimen evidencias y argumentos en contra que de que los prototipos puedan componer (Fodor y Lepore 1996; Connolly *et al.* 2007; Gleitman *et al.* 2012; Machery y Lederer 2012), y quienes sostienen que ambos –prototipos y composicionalidad– sí son compatibles o que, en caso de no serlo completamente eso no supone una limitación explicativa para la teoría (Prinz 2002, 2012; Jönsson y Hampton 2007; Schurz 2012; Del Pinal 2016).

— *Dificultad para explicar cómo se determinan las propiedades relevantes*: ésta es, junto con el problema de composicionalidad, uno de los dos problemas más graves a los que se enfrenta la teoría de prototipos. La cuestión hunde sus raíces en la noción de similaridad, y el hecho de que ésta se evalúe con respecto a un conjunto concreto de propiedades. El problema es que, en ausencia de un conjunto de principios que determinen cuáles son las propiedades relevantes –así como, cuál es su importancia relativa–, no puede determinarse la similaridad entre ningún par de entidades, y los modelos basados en similaridades resultan inútiles[22]. De hecho, conforme indica Medin (1989, p. 1474), buena parte del trabajo explicativo en la teoría es desempeñado, no tanto por la noción de similaridad, como sí por los princi-

[21] En este caso, lo que predice la teoría es que la extensión borrosa del concepto MANZANA-RAYADA satisface la *regla del mínimo* (Zadeh 1965), consistente en que los objetos que pertenezcan a esa extensión han de cumplir en un cierto mínimo grado el ser tanto una cosa –MANZANA– como la otra –RAYADA–. El problema es que un muy buen ejemplo de MANZANA-RAYADA será un mal ejemplo de MANZANA, lo que viola la regla del mínimo antes mencionada.

[22] Por lo tanto, ésta es una dificultad no solo de la teoría de prototipos, sino de todo enfoque basado en similaridades y, en consecuencia, también de la teoría de ejemplares.

pios que guían la selección de propiedades[23]. Sin esos principios las aproximaciones basadas en similaridades se enfrentan a lo que algunos llaman (Machery 2009) –y yo referiré en este trabajo como– el *problema de selección*[24]. Los enfoques basados en similaridades acostumbran a responder a esta cuestión indicando que las propiedades relevantes pueden inferirse del contexto (Torgerson 1965) –entendido como el conjunto de objetos considerados–, sobre la base de su mayor *validez* (Rosch *et al.* 1976; Rosch 1978) o *diagnosticidad* (Tversky 1977; Tversky y Gati 1978; Goldstone *et al.* 1997).

Con respecto a estos dos criterios, la *validez* –o *cue validity*– de una propiedad p, como indicador de una cierta categoría C, no es más que la probabilidad condicionada $P(C|p)$ –es decir, la probabilidad de que un objeto o pertenezca a la categoría C sabiendo que o tiene la propiedad p–. Dada la forma de la probabilidad condicional, a saber, $P(x|y) = P(x \cap y)/P(y)$, la validez de una cierta propiedad p aumenta conforme lo hace la frecuencia con la que la propiedad p aparece juntamente con la categoría C –esto es, $P(C \cap p)$–, y disminuye conforme aumenta la frecuencia con la que la propiedad p aparece asociada con otras categorías distintas de C –pues eso aumenta $P(p)$ sin que se incremente $P(C \cap p)$–. O, dicho de otro modo, una propiedad p será tanto más válida como indicadora de la categoría C cuanto más esté presente en los miembros de C, y cuanto más específica –o diferencial– sea de dichos miembros (esto último equivale a decir que cuanto menos esté presente esa propiedad en no-miembros de C)[25].

[23] Por ello, si el criterio para determinar cuáles son las propiedades relevantes es que tales propiedades intervengan como entradas de sus procesos de categorización asociados (Smith y Medin 1981, p. 16), entonces la explicación de dichos procesos de categorización en términos de meras similaridades resulta circular.

[24] El *problema de selección* continúa el camino marcado por la séptima objeción de Goodman a la noción de similaridad (a saber, que "[l]a similaridad no puede se puede equiparar con, o medir en términos de, la posesión de unas propiedades en común" (Goodman 1972, p. 443)). Recordemos que Goodman basaba su argumento en los tres modos en que puede responderse a la cuestión "¿cuándo son dos cosas similares?": (i) cuando tienen al menos una propiedad en común, en cuyo caso cualquier cosa sería similar a cualquier otra, pues todas tienen algo en común; (ii) solo cuando tienen todas sus propiedades en común, en cuyo caso nada sería similar a nada, pues ningún par de cosas tienen todas sus propiedades en común; y (iii) cuando tienen propiedades importantes en común, lo que conduce directamente a la cuestión de cuáles son tales propiedades importantes y, por consiguiente, al problema de selección.

[25] Sobre esta base puede definirse la validez de una categoría al completo como la suma de las valideces de todas y cada una de las propiedades atribuidas a dicha categoría. Dado que la validez de una categoría no es una probabilidad, podrá tomar valores superiores a 1. En todo caso, una categoría con un valor alto de validez se distinguirá más que categorías que tengan un valor de validez menor. Esto es lo que lleva a Rosch a sostener que, si las *categorías básicas* (como MESA O MANZANA) son las más inclusivas –en el sentido de que sus propiedades están presentes en la mayoría de sus miembros–, entonces son también las que presentan una mayor validez, frente a las *categorías superordinadas* (como MOBILIARIO O FRUTA) cuyos miembros comparten menos propiedades, y también frente a las *categorías subordinadas* (como MESA-DE-COCINA O MANZANA-REINETA) muchas de cuyas propiedades se solapan con las de otras categorías (Rosch *et al.* 1976, p. 384).

Por su parte, Tversky (1977, p. 342) define la *diagnosticidad* de una propiedad[26] como su importancia a efectos de clasificación –o significancia clasificatoria–, determinada ésta como la prevalencia de las clasificaciones realizadas a partir de esa propiedad. De este modo, si una propiedad estuviera presente en todos los objetos considerados, entonces dicha propiedad carecería de diagnosticidad, en la medida en que no aporta nada a la hora de clasificar tales objetos en categorías. Por ejemplo, la propiedad *real* tendría diagnosticidad cero en el contexto de las personas existentes, pues está presente en todos sus miembros; sin embargo, sí tendría un valor de diagnosticidad si el universo considerado se extendiera a las personas de ficción.

Volviendo a la cuestión del problema de selección, éste presenta otra dificultad derivada, asociada con la cuestión de si las creencias que mantienen distintos sujetos acerca de una misma realidad suponen diferentes conceptos de dicha realidad por el mero hecho de distinguirse en algunos de sus elementos (aún cuando la mayor parte de tales perspectivas sea compartida). Por ejemplo, si las creencias que dos sujetos tienen acerca de la categoría HOMBRE difieren en la atribución, o no, de alma inmaterial a sus miembros, ¿tienen ambos sujetos el mismo concepto HOMBRE? Esta cuestión ha sido tradicionalmente referida como el *problema de la estabilidad* de los conceptos (Rey 1983; Smith *et al.* 1984), y deriva del problema de selección porque, en la medida en que no existen unos principios para determinar cuáles son las propiedades relevantes de los conceptos, siempre podrá haber diferencias entre las creencias que distintos sujetos mantienen acerca de una misma categoría. Finalmente, otra cuestión derivada es la de si los conceptos permanecen estables a través de los cambios que una persona experimenta en sus creencias a lo largo de la vida, y su relación con el debate entre invariantismo y contextualismo presentado en el capítulo 1.

— Otras críticas también realizadas a la teoría de prototipos han sido su dificultad para explicar que, en ocasiones, la variabilidad de los miembros de una determinada categoría es importante –tal y como mostró Rips (1989) en tareas de categorización para los conceptos PIZZA y CUARTO-DE-DÓLAR–; y el hecho de que los prototipos suponen una pérdida de información acerca de ejemplos específicos de las categorías[27].

[26] Tversky presenta la noción de diagnosticidad en el seno de su discusión acerca de qué determina la *prominencia* de una cierta propiedad, cuando la concibe en términos de dos componentes, a saber, *intensidad* y *diagnosticidad*. (La *intensidad* estaría asociada a aquellos factores que incrementan la relación señal/ruido, como por ejemplo el volumen de un tono o el tamaño de una letra.)

[27] No obstante, conforme se indica en la nota al pie 33 de este capítulo, tal pérdida sería subsanable si el sistema cognitivo también almacenase –además de la localización del prototipo– información sobre los principales ejemplares observados de cada categoría, los cuales serían empleados, no en tareas de inferencia o categorización, sino únicamente por ciertos procesos de ajuste conceptual.

2.3.2 Teoría de ejemplares

Poco tiempo después de la aparición de la teoría de prototipos surge un segundo enfoque basado en similaridades, conocida como *teoría de ejemplares* (Brooks 1978; Medin y Schaffer 1978). Frente a su teoría "hermana" –la de prototipos–, la teoría de ejemplares rechaza que exista una única representación que resuma toda la información sobre una cierta categoría y, en vez de ello, sostiene que las categorías se encuentran representadas por medio de ejemplos. Bajo esta aproximación, los conceptos serían conjuntos de ejemplares o, en otros términos, cuerpos de conocimiento sobre las propiedades de los miembros individuales de cada categoría. Así, mientras que la teoría clásica representaría el concepto PÁJARO mediante una definición como «animal con pico que vuela y pone huevos», y la teoría de prototipos lo representaría como *algo similar a un gorrión*, la teoría de ejemplares lo haría por medio de un conjunto de ejemplos de esa categoría, tales como { *gorrión*, *jilguero*, *águila*, *gallina*, *avestruz*, etc. }.

En este caso la clasificación de algo bajo una categoría se determinaría calculando su similaridad con respecto a la información almacenada sobre los ejemplos individuales previamente encontrados de cada categoría relevante, y asignándolo a aquella cuya similaridad total sea mayor[28]. De este modo, en la teoría de ejemplares algo sería clasificado como PERRO no por su semejanza a la tendencia central de todos los perros previamente observados –representada bajo la forma de un prototipo–, sino por la suma total de sus semejanzas con todos los perros encontrados. Así, por ejemplo, el modelo de contexto generalizado de Nosofsky (1988a; 2011) define la probabilidad de que un estímulo (o ejemplar) *a* sea clasificado bajo una categoría *I* –esto es, $P(\mathrm{R}_I|a)$ – del modo siguiente[29]:

$$P(\mathrm{R}_I|a) = \frac{\sum_{i \in C_I} n_i s(a,i)}{\sum_K \sum_{k \in C_K} n_k s(a,k)}$$

En donde $s(i,j)$ representa la similaridad entre los estímulos *i* y *j*; la expresión C_I representa al conjunto de ejemplares pertenecientes a la categoría *I*, siendo *K* el conjunto de todas las categorías; y n_i expresa la frecuencia relativa del ejemplar *i*.

Ahora bien, no está absolutamente claro en qué debería consistir la representación de un ejemplar (Hampton 1997b), siendo posibles dos aproximaciones claramente diferenciadas, a saber, multi-prototipo y de instancias (Komatsu 1992). En el primer caso –o *aproximación multi-prototipo*– los ejemplares serían representaciones que abstraen los

[28] Estas dos asunciones (a saber, que la única información almacenada en la memoria son las tendencias centrales asociadas a cada categoría –en el caso de la teoría de prototipos–, y que nuestra memoria recuerda a todos los miembros de cada categoría encontrados en el pasado–para el caso de la teoría de ejemplares–) constituyen dos requisitos muy distintos, tanto en términos de espacio de almacenamiento en la memoria, como de capacidad de procesamiento necesaria para computar las similaridades en tareas de categorización, inferencia, etc.

[29] Obviamente, esta fórmula presenta el problema de que para su evaluación hay que determinar la similaridad entre el estímulo *a* considerado, y todos los ejemplares de todas las categorías almacenadas por la mente del sujeto.

atributos (bajo la forma de tendencias centrales) de distintos grupos de instancias particulares de una categoría, de modo similar a lo que ocurría en el caso de los prototipos[30]. En este caso, el concepto PERRO consistiría en un conjunto de representaciones individuales asociadas a los prototipos de *galgos*, *pastores*, *mastines*, *sabuesos*, *caniches*, *terriers*, etc. En el segundo caso –o *aproximación de instancias*– los ejemplares no abstraerían información sobre los miembros individuales de una categoría, sino que se corresponderían directamente con las instancias individuales. Bajo esta aproximación, el concepto PERRO consistiría en el conjunto de representaciones de todos los perros concretos encontrados por el sujeto (por ejemplo, su propio perro, el perro de su vecino, el perro de su hermano, etc.). De estos dos enfoques ha sido el segundo el que tradicionalmente ha recibido más atención, y con el que habitualmente se identifica la teoría de ejemplares, razón por la cual ésa será la aproximación a la que por defecto me estaré refiriendo en este trabajo. Finalmente, aún dentro de la aproximación de instancias caben diferentes asunciones con respecto a cuáles y cuántos ejemplares de cada categoría se almacenan, a saber: todos ellos, la mayoría de ellos, los más frecuentes, o los más típicos[31].

Entre las ventajas de la teoría de ejemplares destaca el que, siendo una aproximación basada en similaridades –como lo era la teoría de prototipos–, permite dar cuenta de fenómenos no explicados por ésta, como el que los sujetos categorizan más rápidamente objetos similares a otros anteriormente encontrados que objetos nuevos[32] (Medin y Schaffer 1978). Además, la teoría de ejemplares resulta consistente con el hecho de que la precisión en tareas de clasificación mejore cuando aumenta el tamaño de las categorías (Homa *et al.* 1981; Hintzman 1986). En último lugar, la teoría de ejemplares, al igual que la teoría de prototipos, también puede explicar los fenómenos de tipicalidad identificados en muchos conceptos. Sin embargo, en este caso existe una importante diferencia entre ambas pues, mientras que la teoría de prototipos consideraba que la abstracción identificadora de tendencias centrales tiene lugar en el momento de la formación del concepto (prototipo, en este caso), en cambio para la teoría de ejemplares esa abstracción entre instancias ocurriría, no en el momento de la adquisición del concepto, sino cuando ese concepto es empleado por procesos cognitivos que realicen juicios de tipicalidad o hagan uso de ellos[33].

[30] En esta misma línea, varios han sido los autores que han sugerido que prototipos y ejemplares pueden colapsar en un modelo único, en el cual los prototipos fuesen empleados de manera general en tareas de categorización, salvo en situaciones con pocos ejemplares (Knapp y Anderson 1984) o con casos atípicos (Love *et al.* 2004), en los que la categorización tendría lugar sobre la base de ejemplares.

[31] Todo esto contribuyó a que, además del *modelo de contexto* inicialmente propuesto por Medin y Schaffer (1978), pronto surgieran muchas otras propuestas con modelos alternativos que podrían articular la teoría de ejemplares (Brooks 1978; Hintzman y Ludlam 1980; Nosofsky 1984, 1986, 1992a; Estes 1986a, b; Hintzman 1986; Medin 1986; Kruschke 1992; Nosofsky y Palmeri 1997; Lamberts 1998, 2000; Nosofsky y Zaki 2002).

[32] El menor tiempo de procesamiento de objetos previamente vistos –frente a objetos nuevos– podría explicarse asumiendo que los sujetos recuerdan categorizaciones pasadas de objetos concretos, para los cuales no precisan rehacer el proceso de clasificación para cada nueva ocasión.

[33] Esta diferencia es significativa, pues convierte a la teoría de ejemplares en mucho menos eficiente desde un punto de vista computacional –tanto en términos de almacenamiento, como de procesamiento de la información almacenada– que la teoría de prototipos. En cuanto al *almacenamiento*, ambas teorías

En cuanto a sus problemas, la teoría de ejemplares adolece del mismo *problema de selección* que la teoría de prototipos (Machery 2009; Bloch-Mullins 2017), puesto que ambas son aproximaciones basadas en similaridades y ninguna es capaz de explicar cómo se determinan las propiedades relevantes de las categorías. Además, otra dificultad que comparten ambas teorías es el *problema de composicionalidad*, no estando tampoco en este caso claro cómo la teoría de ejemplares podría dar cuenta de nuestra capacidad para clasificar nuevas combinaciones de conceptos (Hampton 1997b).

Por otro lado, la teoría de ejemplares ha sido criticada porque su mayor informatividad (frente a la teoría de prototipos) se logra a costa de la introducción de un significativo problema de economía y coherencia en la teoría. Con respecto al primero –*problema de economía*–, si toda instancia encontrada se almacenase dentro de su respectivo concepto, entonces no habría ningún tipo de economía cognitiva, tanto en términos de almacenamiento como de procesamiento[34] de esa información almacenada. La alternativa es considerar que no todo ejemplar sea almacenado, pero eso requiere que determinadas instancias, bien no se registren, bien se olviden con el paso del tiempo, con el problema de que la teoría no proporciona un modelo sistemático de cómo esto podría ocurrir (Komatsu 1992). En cuanto al segundo –o *problema de coherencia*–, la teoría no explica cómo se forman las categorías, y no aclara tampoco por qué personas con experiencias cotidianas diferentes no acaban teniendo sistemas conceptuales significativamente distintos (Medin 1989; Hampton 1997b).

realizan asunciones muy distintas con respecto a los requisitos de memoria. Por un lado, la teoría de ejemplares considera que nos formamos recuerdos de todos los miembros de una cierta categoría, para emplearlos después en tareas de categorización, inferencia, etc. Por el otro, la teoría de prototipos sostiene que la memoria de largo plazo almacena solamente el prototipo asociado a cada categoría, para su posterior uso en dichos procesos cognitivos. Con respecto al *procesamiento*, la teoría de ejemplares requiere que se compute la similaridad entre cada uno de los miembros del conjunto de ejemplares recuperados de la memoria y el objeto evaluado (es decir, requiere de un cálculo de similaridad por cada ejemplar recuperado); mientras que en el caso de la teoría de prototipos basta con el cálculo de una única similaridad, a saber, entre el prototipo de la categoría y el objeto evaluado.

No obstante, constituye un error pensar que un modelo de prototipos almacena sobre cada categoría –en la memoria de largo plazo– únicamente la localización del prototipo asociado a esa categoría. La razón es que si solo el prototipo fuera registrado en memoria, entonces muchos procesos de modificación de la estructura conceptual de un sujeto no podrían llevarse a cabo. Obviamente, el reajuste de la localización del prototipo como resultado de la exposición a una nueva instancia de esa categoría siempre sería posible (si el sistema cognitivo hubiera almacenado el número de ejemplos previos de esa categoría a los que ha estado expuesto el sujeto). Sin embargo, transformaciones más generales del sistema conceptual asociadas, por ejemplo, al cambio de ejemplares de una a otra categoría, o a la producción de conceptos nuevos –o más específicos que los anteriores–, no serían posibles sin las localizaciones de los principales ejemplos de cada categoría que fueron empleados en el pasado por el sujeto para formarse sus prototipos asociados. Tales ejemplares no serían recuperados a efectos de categorización (por lo que en términos de procesamiento la teoría de prototipos seguiría siendo más eficiente que la de ejemplares), pero sí en algunos procesos de reajuste conceptual.

[34] En este caso, la teoría de ejemplares tendría que explicar cómo es posible que, siendo los conceptos meros conjuntos de instancias, los sujetos utilicen de manera habitual información sobre las tendencias centrales de las categorías.

Otra limitación de la teoría de ejemplares es su reducida capacidad –salvo excepciones muy concretas como, por ejemplo, el modelo de Kruschke (1992)– para atribuir pesos diferenciados a las distintas propiedades de los conceptos (Hampton y Passanisi 2016), lo que constituye una dificultad a la hora de explicar por qué las personas acostumbran a ponderar diferentemente sus distintas propiedades (a la luz de las diferentes importancias que las atribuyen).

Finalmente, también se ha argumentado que, en la medida en que los ejemplares no dan cuenta de un amplio número de fenómenos en psicología de los conceptos (tales como su estructura jerárquica, la inducción o la representación y almacenamiento del conocimiento), no existiría –en realidad– una teoría de los conceptos como ejemplares, sino que lo único que hay es una teoría de categorizaciones basada en ejemplares (Murphy 2016).

2.4. Enfoques basados en explicaciones

De modo análogo a lo ocurrido en el caso de las aproximaciones basadas en similaridades –a saber, que surgieron por la incapacidad de la teoría clásica para explicar los efectos de tipicalidad–, años después surgirán, con objeto de acomodar fenómenos aparentemente no explicados por las teorías de similaridad[35], los *enfoques basados en explicaciones*, también llamados *enfoques basados en conocimiento* (Medin 1989), *enfoques basados en teorías* o, simplemente, *teoría-teoría*.

2.4.1 Teoría-teoría

Una de las primeras propuestas fue la de Murphy y Medin (1985)[36], para quienes los conceptos deberían concebirse en términos de conocimiento teórico o, cuando menos, deberían encontrarse incorporados en el seno de una teoría sobre el mundo. Sobre esta base, las categorías estarían determinadas por unos principios –explicativos y causales– comunes a todos los elementos de cada categoría, principios que determinarán qué correlaciones entre atributos son relevantes para los miembros de esa categoría. Por lo tanto, la categorización de algo bajo un cierto concepto dependería de que los atributos de ese algo encajen con los principios que ese concepto exige a los atributos de sus miembros. O, dicho de otro modo, la clasificación algo bajo un concepto no se limita a una mera comprobación de –o chequeo con– las propiedades asociadas a ese concepto, sino que requiere que el objeto considerado tenga la relación explicativa correcta con la "teoría" que actúa como marco organizador del concepto en cuestión (Medin 1989). Así, bajo

[35] Obsérvese, por ejemplo, la dificultad que tienen las aproximaciones basadas en similaridades para representar relaciones causales y explicativas, en la medida en que dichos enfoques se limitan –por lo general– a representar atributos, sin codificar información con respecto a cómo tales atributos coocurren entre sí (Prinz 2002, p. 77).

[36] En ese mismo año Carey (1985) publica su propio planteamiento, desarrollado independientemente del de Murphy y Medin, aunque en este caso centrado en el ámbito de la psicología del desarrollo. Otros relevantes trabajos pioneros fueron los de Lakoff (1987), Keil (1989) y Gopnik y Wellman (1994).

este enfoque, la categoría SALTAR-VESTIDO-A-LA-PISCINA, aún no estando directamente relacionada con el concepto EBRIO, podría llevar a clasificar a alguien como tal en la medida en que la propiedad EBRIO, en las circunstancias adecuadas, puede actuar como principio explicativo de que alguien haya saltado vestido a la piscina.

No obstante, los principios de la teoría-teoría están mucho menos definidos que los de los enfoques basados en definiciones o similaridades, en la medida en que se pueden concebir de maneras muy diferentes (Laurence y Margolis 1999; Prinz 2002; Machery 2009). Por un lado, cabe considerar que los conceptos son *teorías mentales*, en el sentido de conjuntos de estructuras causales que determinan el conocimiento que el sujeto tiene acerca de una cierta categoría. En este caso, algo cae bajo un concepto cuando sus atributos son –o pueden ser– el resultado de la estructura causal que caracteriza a los miembros de esa categoría (Rehder y Hastie 2001; Rehder 2003a, b). Por otro lado, se puede sostener también que los conceptos son *términos teóricos* (esto es, elementos de las teorías y no teorías mentales completas), determinados / afectados por el papel que esos conceptos juegan en sus correspondientes teorías (Murphy y Medin 1985; Gopnik y Meltzoff 1997), lo que encuentra mejor encaje con la hipótesis de que los conceptos son los constituyentes de los pensamientos. En este segundo caso, los conceptos se relacionarían entre sí del mismo modo a cómo lo hacen los términos de una teoría científica, y su contenido teórico estaría determinado por el papel que cada uno cumple en su teoría asociada.

Sin embargo, todas las aproximaciones a la teoría-teoría comparten la tesis de que los conceptos almacenan conocimiento explicativo de tipo diverso –a saber, nomológico, causal, funcional o genérico– sobre los atributos de los miembros de las categorías. Ese conocimiento contendría información sobre cómo los objetos pertenecientes a cada categoría interactúan con los miembros de otras categorías, a partir del cual podrían realizarse predicciones, explicaciones e interpretaciones[37]. En ambos casos (esto es, con independencia de que se considere que los conceptos son teorías o términos teóricos), sus teorías mentales asociadas constituirían esquemas explicativos de clasificación que permitirían determinar cuándo algo cae bajo una cierta categoría. No obstante, aún a pesar de la función explicativa y predictiva atribuida a las teorías, éstas no tendrán que ser siempre sofisticadas, pudiendo consistir meramente en tener la impresión general de que hay algo causalmente relevante en una categoría que ocasiona algunas o todas las restantes propiedades comunes a sus miembros (Gelman 2005; Carey 2009).

Las evidencias que respaldan este tipo de enfoques surgen de la observación de que en ciertos casos los sujetos categorizan e infieren de modos incompatibles con los principios de los enfoques basados en similaridades. Por un lado, en muchas ocasiones los niños hacen inferencias inductivas sobre la base de pertenencias a categorías, aún cuando dichas inferencias entran en claro conflicto con las apariencias perceptuales como, por ejemplo, cuando consideran que los tiburones respiran como los peces tropicales (y no como los delfines) en virtud de que ambos –tiburones y peces tropicales– son peces, aún cuando su

[37] Algunos autores describen el modo en que se podría adquirir y almacenar ese conocimiento causal en términos de *redes bayesianas* (Gopnik y Schulz 2004), como formalismo basado en modelos gráficos capaz de representar las relaciones causales existentes entre un conjunto de variables (Cooper, 1999; Spirtes *et al.* 2000; Glymour 2001) –o, en este caso, entre los objetos y/o eventos evaluados por los sujetos–.

similaridad es menor que la existente entre tiburones y delfines (Gelman y Markman 1986). Por otro lado, algunos atributos parecen mostrar un carácter "esencial" (u obligatorio) en virtud de las relaciones explicativas existentes entre ellos y otros rasgos de sus respectivos conceptos, de normas que los constriñen, etc. Ejemplos de ello es que las personas, aún cuando atribuyen a la propiedad CURVADO una tipicalidad semejante para el caso de los conceptos BOOMERANG y PLÁTANO, consideran que un objeto recto será –mucho más probablemente– un plátano que un boomerang (Medin y Shoben 1988); o que las personas afirmen que, aún cuando un objeto (imaginado) de tres pulgadas de diámetro es más similar a un cuarto de dólar que a una pizza, lo más probable es que sea una pizza y no un cuarto de dólar porque este último debería tener unas dimensiones uniformes (Rips 1989)[38]. En relación a todos estos casos, la ventaja de la teoría-teoría es su capacidad para incorporar la tendencia que parecen mostrar las personas hacia el pensamiento esencialista[39], en el sentido de que la pertenencia de algo a un grupo es, no tanto una cuestión de presentar un cierto conjunto de atributos, como sí de tener una cierta estructura o propiedad interna conectada de manera explicativa y causal con sus restantes propiedades relevantes.

Conforme ocurría en todas las aproximaciones anteriores, la teoría-teoría tampoco está libre de objeciones. En este caso, la teoría-teoría ha sido criticada porque, bajo la asunción de que toda teoría está hecha de conceptos, si tales conceptos se conciben *como teorías* entonces la propuesta es circular –pues las teorías están constituidas por unos conceptos que, a su vez, son concebidos como teorías–, en cuyo caso la teoría-teoría no estaría explicando lo que los conceptos son. Por otro lado, concibiendo los conceptos como *términos teóricos* se evita el problema de circularidad, pero a costa de no explicar cuál es la estructura de los conceptos[40] ni, por consiguiente, qué es lo que dichos conceptos son.

Además, la teoría-teoría comparte también algunas de las dificultades ya presentes en los enfoques previos, tales como: (a) el *problema de la ignorancia y el error*, presente desde el momento en que los sujetos pueden tener teorías incompletas y/o erróneas acerca de los conceptos; (b) el *problema de selección* en la determinación de los rangos para propiedades continuas causalmente relevantes; (c) el *problema de estabilidad*, dado que el contenido de los conceptos no permanecería invariante ante cambios en sus teorías mentales asociadas; y (d) el *problema de composicionalidad* en la explicación de cómo las teorías asociadas a diferentes conceptos pueden combinarse[41].

[38] Y, aunque existen estudios que cuestionan la generalidad de estos experimentos (Malt 1994; Smith y Sloman 1994; Hampton 1995), ninguno de ellos constituye una evidencia decisiva en contra, en la medida en que existen otras explicaciones alternativas a sus resultados (Prinz 2002, p. 85).

[39] Este esencialismo psicológico es lo que nos permitiría distinguir, por ejemplo, entre un *perro de verdad* y un *perro artificial* (o robot-perro), aún cuando ambos fueran externa y funcionalmente idénticos.

[40] Si los conceptos se conciben como términos teóricos se necesita explicar lo que es un término teórico. ¿Es una definición? ¿Un prototipo? ¿Una teoría? En los dos primeros casos –definición y prototipo– se estaría recurriendo a otra teoría sobre la estructura de los conceptos; mientras que en el tercero se apelaría de nuevo a la teoría-teoría, lo que nos devuelve al ya mencionado problema de circularidad.

[41] Aunque la formulación de estos cuatro problemas (a saber, ignorancia-error, selección, estabilidad y composicionalidad) ha sido expresada para el caso de la identificación de los conceptos con *teorías*, to-

Sin embargo, el problema de la ignorancia y el error unido al problema de selección resultan mucho más graves en el caso de la teoría-teoría, en la medida en que sin poder estar seguros de que no ignoramos o estamos equivocados sobre la teoría asociada a un cierto concepto y, no disponiendo de unos criterios que determinen sus propiedades esenciales, queda en el aire la cuestión de qué es lo que convierte a la teoría asociada a un determinado concepto, en una teoría sobre los miembros de su categoría. En otras palabras, la teoría-teoría deja sin explicar cómo los conceptos tienen contenido intencional (Prinz 2002) o, si lo explica, lo hace en términos de esencias circulares. (Esto último es lo que sucedería, por ejemplo, en una teoría para el concepto PATO que identificara a los *patos* como aquellos *animales cuyos padres son patos*.)

2.5. Otras aproximaciones

Como hemos visto en las secciones anteriores, ninguna de la teorías allí presentadas ha sido capaz de proponer una explicación –que no adolezca de problemas significativos– a la formación y aplicación general de conceptos en tareas de categorización, inferencia, etc. El resultado ha sido la aparición de múltiples propuestas alternativas basadas, bien en la combinación de un subconjunto de las teorías anteriores (lo que ha dado lugar a distintas formas de *pluralismo* e *hibridismo*); bien en el rechazo de algunas de sus asunciones generales, tal y como ocurre cuando se considera que la noción de concepto no es un género natural –*eliminativismo*–, o que no pueden existir conceptos complejos constituidos por otros conceptos más simples –*atomismo*–. El propósito de la presente sección es repasar cuáles son las características, virtudes y sombras de todas esas posturas alternativas.

2.5.1 Atomismo conceptual

Aún cuando todas las teorías hasta aquí presentadas discrepaban acerca de la estructura que tienen los conceptos, todas ellas coincidían en aceptar que dichos conceptos deben tener una estructura interna. Tal estructura interna permitiría articular el principio de composicionalidad que, aplicado al caso de los conceptos, explicaría la productividad y sistematicidad del pensamiento. Ahora bien, frente a ellas, los defensores del atomismo conceptual (Fodor 1990, 1998; Margolis 1998; Millikan 1998, 2000) consideran que la mayoría de nuestros conceptos son atómicos, esto es, carecen de estructura interna.

La motivación del atomismo conceptual procede de la dificultad para alcanzar definiciones por parte de las aproximaciones definicionales, y los problemas para explicar la composicionalidad tanto por parte de los enfoques basados en similaridades como de los basados en explicaciones. Siguiendo en esta misma línea, la *amenaza del holismo* es uno de los principales argumentos esgrimidos en contra de la tesis de que los conceptos tienen estructura y que –por ello– están determinados por sus relaciones (de tipo mereológico o inferencial) con otros conceptos. La idea es que, si unos conceptos dependiesen de otros, entonces no estaría claro cómo delimitar qué profundidad en el conjunto de relaciones

dos ellos pueden reformularse en términos análogos si los conceptos fueran concebidos como *términos teóricos*.

conceptuales debería considerarse en la aplicación de cada concepto, hasta el punto de que –potencialmente– podrían aplicar todas ellas[42] (Fodor y Lepore 2002).

Frente a esta amenaza, los partidarios del atomismo sostienen que los conceptos no tienen constituyentes, razón por la cual el contenido conceptual de un concepto no estaría determinado por unos elementos constituyentes y estructura (que aquí se niegan), sino por la información portadora de las relaciones causales entre cada concepto y sus referentes. Bajo este enfoque el contenido conceptual sería mera referencia, en base a lo cual diríamos que alguien posee un cierto concepto si dispone de una entidad mental capaz de establecer las relaciones causales adecuadas entre ese concepto y sus referentes. O, dicho de otro modo, el significado –o identidad– de un concepto no estaría determinado por sus relaciones con oros conceptos, sino por las relaciones causales que ese concepto mantenga con las cosas del mundo, las cuales actuarían como conexión (en el sentido de correlación fiable) entre tal concepto y las propiedades que representa. Así, por ejemplo, el concepto CABALLO no estaría en relación con otros conceptos, tales como ANIMAL, PEZUÑAS, etc., sino que expresaría la propiedad CABALLO mediante una ley causal que uniría dicho concepto con la propiedad *ser un caballo*, que es lo que garantizaría la existencia de una relación correcta entre la mente y el mundo. Por consiguiente, en esta propuesta desaparece el problema de la ignorancia y el error, dado que no importa lo que se crea sobre los caballos, en tanto en cuanto el concepto CABALLO y la realidad del mundo *caballo* estén conectadas del modo adecuado.

Finalmente, el atomismo conceptual también ha recibido objeciones, principalmente debidas a su *debilidad explicativa* cuando –bajo la asunción de que los conceptos no tienen estructura– se intenta dar cuenta de: (i) fenómenos psicológicos tales como categorizaciones o inferencias; (ii) cómo conceptos no-atómicos –esto es, complejos– pueden componer; (iii) las intuiciones que tienen los sujetos acerca de que ciertas relaciones entre conceptos son analíticas; y (iv) los ya mencionados fenómenos de tipicalidad. El problema en todos estos casos es que sus fenómenos asociados son difíciles de explicar en ausencia de relaciones entre conceptos, que es justamente lo que los atomistas conceptuales rechazan que exista.

No obstante, el mayor problema del atomismo es la cuestión de si conlleva la *aceptación del nativismo radical* y, por consiguiente, las dificultades asociadas a esta postura ya vistas en el capítulo anterior. Ésa es la conclusión que se deriva de los argumentos nativistas contra la tesis de que los conceptos primitivos –y, por tanto, atómicos– puedan ser aprendidos (Fodor 1975, 1980a, 1981a; Jackendoff 1989; Carey 2009), el cual se podría resumir del modo siguiente:

(1) Todo mecanismo de aprendizaje se reduce a la formación y testeo de hipótesis.
(2) Las hipótesis que juegan un papel en la adquisición de un nuevo concepto deben formularse en términos de los conceptos disponibles y el principio de composicionalidad.

[42] Una posible respuesta a este argumento es la de adoptar un planteamiento de tipo localista, conforme al cual únicamente serían relevantes algunas –que no todas– de las relaciones mantenidas por los conceptos (Weiskopf 2009b).

(3) Los conceptos primitivos no pueden expresarse en términos de otros conceptos.
(4) Luego los conceptos primitivos no pueden aprenderse y, por tanto, son innatos.

Con el problema de que quien acepte la anterior conclusión tendrá que enfrentarse a la implausible consecuencia de que son innatos también candidatos tan improbables como los conceptos de XILÓFONO, HÉLICE O ELECTRÓN.

En todo caso, no todos los defensores del atomismo aceptan que el carácter atómico de los conceptos conduce necesariamente a que éstos deban ser innatos. Así, por ejemplo, Margolis (1998) argumenta a favor de la tesis de que los conceptos pueden ser –a la vez– atómicos y aprendidos, en cuyo caso el atomismo no tendría que enfrentarse a los problemas que afrontan los nativistas radicales.

2.5.2 Teorías híbridas

Una de las propuestas surgidas como reacción al problema de que ninguna de las principales teorías sobre la estructura de los conceptos sea capaz de explicar de modo general los diversos fenómenos cognitivos asociados a la noción de concepto, han sido las *teorías híbridas* –también llamadas *hibridismo conceptual*– (Smith *et al.* 1974; Osherson y Smith 1981; Nosofsky *et al.* 1994; Keil *et al.* 1998; Anderson y Betz 2001; Vicente y Martínez-Manrique 2016). Frente a las propuestas "puras", los defensores del hibridismo conceptual sostienen que, aún cuando cada categoría estaría representada por un único concepto (lo que diferencia a las teorías híbridas de las posturas de tipo pluralista[43]), tales conceptos estarían divididos en partes, las cuales tendrían asociadas distintas clases de conocimiento, articuladas mediante estructuras conceptuales diferentes.

No obstante, existen múltiples versiones de hibridismo, en función de qué elementos se entienda que constituyen los conceptos, y cómo se piense que esos elementos se conectan y coordinan. Por un lado, algunos consideran que los conceptos están constituidos por un *procedimiento de identificación*, basado en similaridades y empleado para clasificaciones rápidas, y un *núcleo* de tipo definicional que intervendría cuando razonamos sobre los conceptos (Smith *et al.* 1974; Miller y Johnson-Laird 1976; Osherson y Smith 1981; Smith y Medin 1981; Landau 1982)[44]. Otros sugieren, en cambio, que los conceptos pueden combinar una parte basada en reglas con excepciones basadas en ejemplares (Nosofsky *et al.* 1994; Anderson y Betz 2001). Otra posibilidad –no incompatible con las anteriores– es que los conceptos combinen los dos enfoques basados en similaridades, a saber, una parte prototípica con las tendencias centrales encargada de la identificación de nuevas instancias no-atípicas (lo cual sería útil en términos de economía y coherencia), y otra parte que almacenara ejemplares particulares de esa categoría empleados en etapas tempranas o intermedias del proceso de aprendizaje (Homa *et al.* 1991). Finalmente, otra

[43] Esto no es óbice para que no puedan articularse propuestas que combinen hibridismo y pluralismo (Laurence y Margolis 1999; Margolis y Laurence 2010; Rice 2016).

[44] De modo similar Neimark (1983) distingue –para el caso de las clasificaciones– entre modelos de competencia (de tipo clásico), que actuarían cuando se dispone de un algoritmo clasificador, y modelos de rendimiento (basados en prototipos), como aproximación heurística por defecto en ausencia de reglas de clasificación. Por su parte, Rosch (1983) caracteriza al razonamiento lógico de modo clásico, frente al razonamiento analógico por puntos de referencia, que es concebido en términos de prototipos.

opción a considerar es la integración de un enfoque basado en similaridades con otro basado en explicaciones (Medin 1989) lo cual, de nuevo, es perfectamente compatible con todas las aproximaciones antes mencionadas.

Ahora bien, las teorías híbridas son en ocasiones criticadas porque las clasificaciones que resultan de las diferentes partes que, supuestamente, constituyen los conceptos pueden ser inconsistentes entre sí. De hecho, eso es a lo que apuntan los estudios de Malt (1994), cuando las personas categorizan un líquido como AGUA atendiendo, no solo a su composición química (porcentaje de H_2O) –característica de su concepción esencialista/definicional–, sino también a su origen, uso y localización. En este caso el problema estriba en que los juicios sobre la concentración de H_2O en un líquido no predicen bien si ese líquido será considerado AGUA o no, razón por la cual sus clasificaciones (basadas, bien en definiciones, bien en similaridades) en muchas ocasiones se contraponen entre sí. Por ejemplo, el líquido de una taza de té puede ser considerado un ejemplo de NO-AGUA, aún cuando el porcentaje de agua en él –un 91%– sea mucho mayor que en otros líquidos que sí son considerados AGUA, conforme ocurre en los casos del agua de una piscina o del agua de mar –cuyas concentraciones de H_2O son del 81% y 78,7%, respectivamente–.

Por otro lado, las teorías híbridas también han sido criticadas porque la condición de coordinación entre las diferentes partes de un concepto parece contradecir el hecho de que muchos usos de las palabras asociadas a esos conceptos resultan ambiguos. Así, Machery (2009) sostiene que las distintas partes de un concepto no solo deberían estar *conectadas* (de manera que una clasificación positiva por parte de cualquiera de ellas hiciera que el resto estuviera disponible para otros procesos cognitivos), sino también *coordinadas* entre sí (para que las distintas partes de un concepto no generen resultados –por ejemplo, categorizaciones– inconsistentes). No obstante, argumenta Machery, el hecho de que afirmaciones como "Tina Turner es una abuela" tengan dos lecturas –una verdadera y otra falsa–, en función de si el término *abuela* se aplica en un sentido definicional o prototípico[45], podría ser indicativo de que la coordinación entre las distintas partes de un concepto no opera del modo que necesitan los defensores del hibridismo.

Finalmente, el hibridismo también ha sido criticado por los defensores del pluralismo conceptual (Weiskopf 2009a), bien por asumir que lo que interviene en toda tarea cognitiva es el concepto "completo" –o unión de todas las diversas partes asumidas por el hibridista, esto es, prototipos, ejemplares, teorías, etc.– por la inflación representacional que eso supone; bien, si únicamente se emplearan algunas partes del concepto en cada ocasión, por no explicar cómo tales partes relevantes se seleccionan. En el primer caso la objeción consiste en que los conceptos "completos" son demasiado grandes para su empleo por la memoria de trabajo. En el segundo, el hibridismo es criticado por no proporcionar un criterio que individúe qué información de la categoría será empleada en cada caso, sin caer al hacerlo en planteamientos de corte pluralista[46].

[45] Este mismo problema puede ser extrapolado a otros casos similares de ambigüedad o polisemia, con combinaciones de estructuras conceptuales diferentes en donde, por ejemplo, los conceptos tuvieran una parte basada en teorías y otra en prototipos (Machery y Seppälä 2011).

[46] Ambas objeciones han recibido respuestas desde el ámbito hibridista. Con respecto a la primera se ha indicado que los constituyentes de los pensamientos presentes en la memoria de trabajo no serían los

2.5.3 Teorías pluralistas

Las *teorías pluralistas* –o *pluralismo conceptual*– constituyen una segunda posibilidad a la hora de defender que los conceptos pueden tener múltiples estructuras asociadas. En este caso, esas diversas estructuras no deberían identificarse con distintas partes de un mismo concepto –conforme sostiene el hibridismo–, sino con diferentes tipos de conceptos que operarían de modo específico en cada tarea o competencia cognitiva.

La idea subyacente a este planteamiento es que los conceptos no constituyen un género natural único, sino se dividen en géneros distintos, en la medida en que no existe ninguna representación individual que pueda explicar todos los fenómenos empíricos de los que los conceptos son responsables (Piccinini y Scott 2006; Weiskopf 2009a)[47].

Y, aunque similares, pluralismo e hibridismo son propuestas diferenciadas. Por un lado, el hibridismo consideraría que el concepto asociado a una determinada categoría –por ejemplo, la categoría *perro*– puede tener asociadas diversas estructuras conceptuales (a saber, definiciones, prototipos, ejemplares o teorías), no siendo éstas conceptos distintos, sino las partes constitutivas de un único concepto PERRO. En cambio, para el pluralismo no hay tal cosa como el concepto PERRO, sino muchos conceptos-perro, cada uno de ellos con una estructura conceptual diferente. Así, por ejemplo, podría haber un concepto PERRO_1 –con estructura prototípica– que explicase algunos casos de categorización e inferencia, otro concepto PERRO_2 –basado en ejemplares– que diera cuenta de otras categorizaciones e inferencias diferentes, y un tercer concepto PERRO_3 –de tipo definicional– que explicara cómo razonamos sobre los perros.

Ahora bien, uno de los retos a los que se enfrenta el pluralismo es el de explicar qué unifica a toda esa pluralidad de conceptos, y los convierte a todos ellos en conceptos de una misma categoría. O, dicho de otro modo, por qué todos esos diferentes conceptos PERRO_i deben ser considerados conceptos de *perro*. Y, aunque una posible respuesta sería decir que aquello que los unifica es que todos refieren a la misma categoría, no está claro que eso resulte evidente, en la medida en que ninguno de ellos fija la misma referencia (siendo ésta precisamente una de las virtudes de las aproximaciones pluralistas).

Sin embargo, posiblemente el mayor problema para el pluralismo es que las mismas consideraciones que conducen a su adopción (a saber, que los conceptos cumplen múltiples funciones cognitivas, y que diferentes tipos de conceptos –con estructuras conceptuales diversas– pueden intervenir en cada una de ellas), pueden llevar a adoptar tesis de corte eliminativista, que consideran que el motivo por el que ninguna estructura conceptual es capaz de explicar todos esos fenómenos cognitivos es porque los conceptos no constituyen un género natural.

conceptos completos, sino versiones reducidas suyas. En cuanto a la segunda, Vicente y Martínez Manrique (2016) consideran una *coactivación funcional estable* permitiría explicar cómo cada concepto se individúa –o, bajo el enfoque de Machery, cómo sus diferentes partes pueden coordinarse–.

[47] Adicionalmente, Machery (2009) distingue dos tipos de pluralismo. Por un lado estaría el *pluralismo de alcance*, para el cual los diversos tipos de concepto están asociados a distintos tipos de entidades, eventos, substancias, etc. Por el otro estaría el *pluralismo de competencia*, cuando los distintos tipos de concepto se asocian con diferentes competencias cognitivas.

2.5.4 Eliminativismo

Desde la perspectiva del *eliminativismo*, si los conceptos fuesen un género natural entonces deberían presentar un conjunto de elementos en común relevantes, susceptibles de ser identificados con métodos empíricos, por lo que el hecho de que esos elementos en común no estén presentes en los diferentes tipos de concepto (asociados a distintas estructuras conceptuales) sería indicativo de que esos conceptos no son un género natural. En consecuencia, la recomendación del eliminativista es que la noción teórica de "concepto" debería ser abandonada (Machery 2005, 2009; Malt 2010), tras lo cual la psicología se debería dejar de dedicar al estudio de los conceptos, pasando a centrarse en el estudio de prototipos, ejemplares, teorías, etc.

No obstante, el eliminativismo ha recibido muy diversas críticas. En primer lugar, algunos consideran que el criterio exigido por Machery a los géneros naturales resulta demasiado exigente, y que deja fuera casos de géneros naturales de utilidad tanto para la psicología (Margolis y Laurence 2010; Prinz 2010) como para la ciencia (Gonnerman y Weinberg 2010)[48]. Otros han sostenido que, aun aceptando el criterio de Machery para los términos de género natural, el eliminativismo conceptual no se sigue de él, puesto que los conceptos podrían ser géneros naturales conforme a ese criterio (Samuels y Ferreira 2010; Weiskopf 2010). Finalmente, algunos autores sugieren que el criterio empleado por Machery es demasiado fuerte, y que –aún a pesar de las críticas de Machery– la noción de concepto, bien en un sentido funcional (Lalumera 2010; Strohminger y Moore 2010), bien como elemento integrador de alto nivel (Hampton 2010), continúa siendo de utilidad en el ámbito de la ciencia cognitiva[49].

2.6. Recapitulación

En el presente capítulo han sido presentadas las principales teorías sobre la estructura de los conceptos, junto con sus más importantes puntos fuertes y débiles. También se ha visto que ninguna de las teorías "puras" –a saber, enfoques basados en definiciones, similares o teorías– era capaz de dar cuenta de todos los fenómenos empíricos relevantes. O, dicho de otro modo, no parece que en la mente opere una única teoría de conceptos, sino que intervienen elementos procedentes de distintas teorías.

Por consiguiente, en mi opinión la solución pasa por una aproximación de tipo híbrido o pluralista pues, como hacen muchos otros autores, me inclino hacia el rechazo de planteamientos eliminativistas como el de Machery, dado que no veo justificada su aceptación en la medida en que (i) la noción de concepto sigue siendo útil en el ámbito de la psicología, y (ii) pluralismo e hibridismo constituyen alternativas muy plausibles al respecto. No obstante, y aunque en principio me inclino más por una aproximación hibri-

[48] Considérese, por ejemplo, el caso de los términos de género natural *módulo*, *computación* o *representación* –en la psicología–, o *algoritmo*, *partícula subatómica* o *nutriente* –en la ciencia–.

[49] Para conocer la respuesta eliminativista a todas estas –y otras– objeciones véase Machery (2010a, 2010b).

dista, en mi trabajo no argumentaré en su favor frente al pluralismo, puesto que las tesis que aquí defenderé resultan compatibles con cualquiera de estos dos enfoques.

Por último, de entre los diferentes enfoques a la estructura de los conceptos, el presente trabajo asumirá de manera general una aproximación basada en prototipos –que podría estar embebida en el seno de una teoría hibridista o pluralista más amplia (por lo que no supone un rechazo de ninguna de las otras teorías alternativas)–. La razón es que, además de sus ya mencionados puntos fuertes (como, por ejemplo, su mayor economía cognitiva en términos de memoria y procesamiento, frente a otras teorías como la de ejemplares), una teoría de prototipos basada en espacios de similaridad permite dar respuesta a uno de los principales problemas del empirismo, a saber, la cuestión de cómo los conceptos primitivos se adquieren sin caer en circularidad –tal y como mostraré en el capítulo 6–.

Finalmente, a pesar de que la noción de prototipo no es crucial –desde un punto de vista teórico– para el desarrollo de las principales tesis defendidas en los capítulos 5 y 6 –en donde la única condición crítica es la asunción de un marco contextualista–, su adopción general como teoría-base en este trabajo facilitará la presentación de los problemas enfrentados, en la medida en que éstos serán ejemplificados y discutidos en primera instancia para una aproximación basada en la teoría de prototipos (aún cuando las conclusiones alcanzadas sean luego generalizadas a marcos más amplios).

This page intentionally left blank

Capítulo 3: Medidas y modelos de similaridad conceptual

From causes, which appear similar, we expect similar effects. –
David Hume (1741, IV ii 20)

For surely there is nothing more basic to thought and language than our sense of similarity; our sorting of things into kinds. –
Willard V. Quine (1969, p. 116)

Conforme indiqué al final del capítulo anterior, en este trabajo asumiré una aproximación a la noción de concepto basada en la teoría de prototipos, y articulada por medio de espacios de similaridad conceptual. En consecuencia, mi enfoque se enmarca dentro de los modelos de similaridad de tipo geométrico. No obstante, los modelos geométricos no son la única forma en que puede caracterizarse la idea de similaridad, razón por la cual el presente capítulo tiene como propósito realizar un repaso de los más relevantes modelos contemporáneos de similaridad, de sus principios y motivaciones, así como de las principales ventajas e inconvenientes de cada uno de ellos. Para ello seguiré la habitual distinción entre modelos geométricos, de rasgos, basados en alineamientos y transformacionales (Goldstone y Son 2005).

Con ese objeto, y tras realizar –en la primera sección– una sucinta introducción de la noción de similaridad, y de las razones de su importancia en filosofía, psicología y ciencia cognitiva, efectuaré un breve repaso –segunda sección– de cuál ha sido la historia de esta noción a lo largo del pasado siglo XX, comenzando con su papel en el *Aufbau* de Carnap (1928) y terminando con las matizaciones y críticas a la misma realizadas por Quine (1969) y Goodman (1951; 1972).

Tras ello presentaré las distintas formas en que puede modelarse un enfoque basado en similaridades, como es la teoría de prototipos[1]. Comenzaré en la sección tercera con la caracterización contemporánea del *modelo geométrico* que, en cierto modo, da continuidad a propuestas como las de Carnap y Quine. Buena parte de dicha sección estará

[1] Aunque mi trabajo estará centrado en la teoría de prototipos, los modelos estudiados en este capítulo son modelos de similaridad *en general*, por lo que serían válidos tanto para la teoría de prototipos como para la teoría de ejemplares, en la medida en que ambas son aproximaciones basadas en similaridades.

dedicada a los tres axiomas métricos asumidos por este tipo de modelo (a saber, minimalidad, simetría y desigualdad triangular), así como a las críticas recibidas por parte de aquellos que consideran que existe evidencia empírica de la violación de esos axiomas, y a los argumentos de quienes sostienen que tales aparentes violaciones se pueden explicar recurriendo a la noción de contexto.

A continuación –sección cuarta– presento el *modelo de rasgos*, como primera propuesta alternativa surgida en respuesta a los problemas del modelo geométrico relativos a la violación de sus axiomas. No obstante, el hecho de que ni el modelo geométrico, ni tampoco el modelo de rasgos, sean capaces de caracterizar adecuadamente cómo los atributos de los objetos se encuentran organizados ha dado lugar a la aparición de otras aproximaciones a la noción de similaridad (a saber, modelos basados en alineamientos y modelos transformacionales) que sí tienen en cuenta esas relaciones estructurales. En la sección quinta presentaré los *modelos de alineamiento*, cuyo principal elemento característico es que no solo evalúan el encaje (en un objeto) de un conjunto de atributos concreto, sino que también comprueban si los atributos presentes están organizados del modo adecuado. Finalmente, en la sección sexta estudiaré el caso de los *modelos transformacionales*, en los que la similaridad entre dos objetos se encuentra determinada por la distancia transformacional –que no geométrica– existente entre sus representaciones.

3.1. Noción de similaridad

Los términos *similaridad*, *similitud* y *semejanza* se emplean indistintamente para referir a la relación existente entre dos entidades cuando éstas no son ni idénticas, ni iguales, ni distintas, sino que tienen a un mismo tiempo algo igual y algo diferente (Ferrater Mora 1994, p. 636). A pesar del carácter aparentemente vago de esta descripción, la similaridad desempeña un papel clave en muchas teorías del conocimiento y comportamiento, en las que actúa como principio organizador mediante el cual los sujetos clasifican los objetos, se forman conceptos y realizan generalizaciones (Tversky 1977, p. 327). En esos casos, la evaluación de la similaridad es fundamental para la cognición, pues revela el mundo cuando lo concibe como un lugar lo suficientemente ordenado como para que objetos y eventos similares (esto es, clases de objetos/eventos[2]) se comporten de modo parecido en sus aspectos importantes.

La idea de similaridad resulta básica para aprender, conocer y pensar, y han sido pocos los epistemólogos empiristas que se han resistido al empleo de esta noción a la hora de explicar cómo los conceptos se forman y aplican. Así, por ejemplo, Quine (1969, pp. 117 y 133) sostenía que tenemos expectativas razonables cuando en circunstancias similares

[2] Esto ha llevado a sostener que las nociones de similaridad y tipo –categoría o género natural– son, en último término, una misma noción (Quine 1969, p. 119). Y, aún cuando Quine no lo consideraba posible, podría intentar argumentarse que cada de ellas es definible en términos de la otra, en la medida en que (i) son similares aquellas cosas que pertenecen a una misma categoría, y (ii) una categoría está constituida por aquellas cosas similares entre sí (o, en términos comparativos, por aquellas cosas más similares entre sí que a cosas pertenecientes a otras categorías). Y, aunque no es mi propósito defender en este trabajo tal identificación, lo que sí resulta poco cuestionable es que ambas nociones se encuentran fuertemente interrelacionadas.

anticipamos que causas parecidas provocarán efectos semejantes, razón por la cual fundamenta predicciones, inferencias y categorizaciones[3], siendo una fuente de información general. Además, la idea de similaridad también ha sido invocada para dar cuenta de la creatividad humana –tanto científica como de otros tipos–, siendo una explicación habitual que los actos de creatividad tienen lugar por medio de analogías entre dominios conocidos y nuevos (esto es, mediante la identificación de elementos similares en ambos dominios). Y, en virtud de esta función unificadora de objetos semejantes, la similaridad ha sido descrita como "la quilla y columna vertebral de nuestro pensamiento" (James 1890, vol. I, p. 459).

Además, la atribución de un papel a la similaridad en funciones cognitivas tales como categorización, inferencia, predicción, resolución de problemas, etc., encuentra respaldo en un amplio número de estudios de psicología experimental. Así, por ejemplo, efectuamos categorizaciones sobre la base de similaridades (Nosofsky 1986; Smith *et al.* 1998) –y lo hacemos incluso cuando disponemos de caracterizaciones detalladas de las categorías (Smith y Sloman 1994)–; generamos nuevas categorías de objetos en función de su similaridad (Nosofsky 1984; Estes 1986b); inferimos entre categorías en base a la similaridad existente entre ellas –razonamiento deductivo– (Sloman 1998), y esas inferencias son más fuertes cuando los ámbitos/dominios de la similaridad y la inferencia coinciden (Heit y Rubinstein 1994); hacemos predicciones sobre casos nuevos en base a su similaridad con otros conocidos –inferencia inductiva– (Smith 1989; Osherson *et al.* 1990; Sloman 1993); aplicamos la similaridad a la toma de decisiones (Smith y Osherson 1989); intentamos resolver problemas similares a otros encontrados en el pasado aplicando principios similares a los que funcionaron en estos últimos (Ross 1987); e incluso hay quien sostiene que toda representación es representación de –que no *por medio de*– similaridades (Edelman 1995, 1998).

Por consiguiente, el estudio de la similaridad resulta necesario para comprender las entidades mentales y los procesos que operan sobre ellas, así como para clarificar el papel jugado por la similaridad (a caballo entre la percepción y la cognición de alto nivel) en la formación de conceptos.

3.2. Breve historia de la similaridad

En este punto, y antes de proceder con el estudio de los principales modelos contemporáneos de similaridad, efectuaré un breve repaso de la historia de esta noción. Aunque las ideas de semejanza y similaridad han desempeñado un papel muy relevante desde el mismo origen de la filosofía –recuérdese, por ejemplo, la semejanza por participación[4] de Platón, o la asociación por semejanza de Aristóteles[5]–, en esta sección me centraré en la historia de la similaridad a lo largo del pasado siglo XX.

[3] Esto se apoya en la asunción de que al aumentar la similaridad entre dos ítems *a* y *b* aumenta la probabilidad de inferir correctamente que *b* presenta la propiedad *F* a partir del hecho de que *a* presente dicha propiedad *F* (Tenenbaum 1999).

[4] Platón, *Parménides* 129a.

[5] Dentro de los tres tipos de asociaciones distinguidas por Aristóteles, a saber, por semejanza, por contigüidad y por contraste (Aristóteles, *Acerca de la memoria y de la reminiscencia*, 451b 18-22).

Así, en el primer cuarto de siglo Carnap, en su intento por reconstruir la estructura completa de las propiedades del mundo sobre la base de una única relación fenoménica[6], consideraba que esa relación básica era el *recuerdo de semejanza*, a partir de la cual se podrían definir todos los otros conceptos y relaciones (Carnap 1928, §78). Allí, en el *Aufbau*, Carnap había previamente definido la *similaridad* como una relación reflexiva y simétrica[7] $s_{\mathrm{d}}(a,b)$ entre dos experiencias elementales[8] –esto es, objetos o eventos– (*ib.*, §11). Sobre esta base Carnap sostenía que dos objetos eran similares en caso de que tuvieran alguna propiedad en común, y no lo serían si no tuvieran ninguna propiedad compartida (*ib.*, §70). A partir de esta definición de similaridad se puede examinar el mundo y determinar qué objetos son similares a qué otros lo que, en la terminología de Carnap, da lugar a la producción de *círculos de similaridad* identificados con las extensiones de las clases de objetos que son similares por pares (es decir, de todos aquellos elementos tales que, tomados de dos en dos, sus pares comparten alguna propiedad).

Ahora bien, la idea de similaridad postulada por Carnap no está libre de problemas, tal y como pronto ilustró su discípulo Quine. Con respecto a esto Quine (1969), tras mostrar que la similaridad entre dos cosas postulada por Carnap equivale a que dichas cosas pertenezcan a una misma clase[9], observó que una noción *binaria* de similaridad[10] –como la de Carnap– presentaba dificultades cuando se consideraban los conjuntos de cosas COLOREADAS y de cosas ROJAS. El problema era que si las cosas COLOREADAS son una clase, entonces todo par de cosas coloreadas serían similares, por lo que el conjunto de cosas ROJAS sería demasiado estrecho como para ser una clase, lo cual resulta inadecuado. De modo alternativo, si las cosas ROJAS fueran una clase entonces no todo par de cosas coloreadas serían similares, por lo que el conjunto de cosas COLOREADAS sería demasiado amplio como para ser una clase, lo cual también es inadecuado[11].

Por otro lado Quine, en su crítica de la similaridad de Carnap, también anticipa el *problema de selección* –luego explicitado en la séptima objeción de Goodman (1972)– cuando señala los problemas de definir la idea de similaridad a partir de la idea de propiedad, lo cual obliga a saber previamente qué es una propiedad. Aquí la dificultad surge cuando se considera cómo puede definirse la idea de propiedad sin recurrir a las nociones

[6] Relación *fenoménica* entendida en el sentido de relación *entre experiencias elementales*.

[7] Frente a la relación de *equivalencia* que –a diferencia de su concepción de la relación de similaridad–, además de reflexiva y simétrica, tiene que ser también transitiva. Y también frente al *recuerdo de semejanza*, el cual era una relación de tipo asimétrico.

[8] En esta sección referiré con el subíndice $_{\mathrm{d}}$ a la noción de *similaridad diádica* de Carnap, y con el subíndice $_{\mathrm{t}}$ a la idea de *similaridad triádica* postulada por Quine.

[9] Esta idea había sido previamente anticipada por Leibniz en A64 107 (1678?) cuando mostraba que las oraciones con la forma "Pedro es similar a Pablo" pueden reducirse a oraciones con la forma "Pedro es A y Pablo es A", lo cual equivale a sostener que dos objetos son similares en caso de que compartan –al menos– una propiedad (esto es, en caso de que caigan bajo un mismo concepto –o clase–).

[10] *Binaria* en el sentido de ser una cuestión que solo admite los valores *sí* y *no*.

[11] Recientemente, Mormann (1994, 2009) y Leitgeb (2007) han desarrollado una formulación matemática precisa de muchos de los problemas planteados por Quine y Goodman, junto con una posible respuesta a los mismos.

de clase o similaridad, lo cual excluye la posibilidad de decir que una propiedad es aquello compartido por un cierto grupo –o clase– de objetos, y también la de decir que una propiedad es aquello compartido por cosas similares. La conclusión de Quine (1969, p. 117) era que definir la idea de similaridad sobre la noción de propiedad equivalía a dejar sin explicar lo que la similaridad es.

La propuesta de Quine consiste en substituir la similaridad diádica –y absoluta– de Carnap por una similaridad $s_t(a,b,c)$ de tipo comparativo, basada en la relación triádica "*a* se parece más a *b* que a *c*", cuya ventaja es que admite relaciones de inclusión entre clases/propiedades. Así, por ejemplo, con una relación de similaridad triádica tanto el conjunto de cosas ROJAS como el de cosas COLOREADAS podrían ser clases, lo cual sería posible si la similaridad entre los elementos del conjunto de cosas ROJAS fuese mayor que su similaridad con respecto al resto de cosas COLOREADAS. La similaridad triádica de Quine rompe –o, cuando menos, desdibuja– la identificación entre las nociones de similaridad y clase[12].

Por su parte Goodman (1972), en su renombrado artículo en contra de la idea de similaridad, planteará siete objeciones en contra de esta noción, siendo las dos últimas (asociadas a la relación entre la similaridad y las propiedades de los objetos) las más interesantes. Con respecto a la objeción sexta, en ella Goodman (*ib.*, p. 441) sostiene que una similaridad diádica entre objetos no es suficiente para definir propiedades. En este caso su punto es que no basta con que los elementos de una clase tengan –dos a dos– al menos una propiedad en común, pues eso no garantiza que exista una propiedad común a todos los elementos de dicha clase[13]. Así, por ejemplo, dados tres discos *a*, *b* y *c* divididos en dos mitades, cada una pintada de un color distinto, y tales que sus colores fuesen rojo-azul (disco *a*), azul-amarillo (disco *b*) y amarillo-rojo (disco *c*), cada par de objetos de este grupo tiene una propiedad en común, aún cuando no hay ninguna propiedad común a

[12] Obsérvese que la noción de similaridad de Carnap y, sobre todo, la de Quine constituyen un claro antecedente de los modelos de tipo geométrico. Por un lado, un *círculo de similaridad* sería, para Carnap (1928, §70, §80 y §111), aquella clase *C* de objetos tales que cada par de elementos en *C* son similares (es decir, comparten alguna propiedad) y ningún elemento fuera de *C* es similar a ningún elemento de dicha clase. Y, aunque el cuasi-análisis de Carnap no se limita a estructuras de similaridad en espacios métricos, el modo de construcción que propone produce –cuando se considera el caso de un espacio geométrico– círculos o esferas de cualidades, esto es, propiedades con forma esférica.

Por su parte, la similaridad comparativa de Quine también produce conjuntos de cualidades esféricas cuando se considera el conjunto de cosas que difieren menos de "algo" con respecto a una cierta norma central (Quine 1969, p. 119). En este caso, bastaría con distinguir un objeto *a* que actúe como *caso paradigmático* –o norma central– de la clase, y otro objeto *b* que actúe como *caso límite*; y luego definir la clase "con paradigma *a* y límite *b*" como el conjunto de aquellas cosas que son más similares a *a* que lo que *a* lo es a *b*, lo cual produciría espacios de cualidades esféricas.

[13] Este problema ya había sido presentado anteriormente por Goodman, y referido como el *problema de la comunidad imperfecta*, en su discusión de las dos principales razones en contra de la posibilidad de que las propiedades puedan definirse mediante la noción de similaridad carnapiana (Goodman 1951, pp. 162-164). La otra dificultad sería el llamado *problema de la compañía*, asociado al hecho de que si las propiedades son conjuntos maximales de objetos –en el sentido de que incluyen a todos los objetos similares en ese aspecto– entonces una propiedad que aplique solo a un subconjunto de los objetos que presenten otra propiedad no podría ser recuperada por medio del cuasi-análisis (*ib.*, pp. 157-161). En todo caso, ambas dificultades habían sido ya identificadas por el propio Carnap (1923) antes de que Goodman las señalara como problemas relevantes para el método del cuasi-análisis.

todos ellos. No obstante, tanto el modelo de rasgos de Tversky como los modelos geométricos contemporáneos parecen capaces de sortear esta dificultad.

En cuanto a su séptima objeción, Goodman (*ib.*, pp. 443-446) sostiene que la similaridad no equivale a, ni puede ser medida en términos de, la posesión de propiedades comunes. En este caso Goodman estudia los distintos modos en que puede definirse la similaridad en términos de propiedades, y concluye que ninguno es satisfactorio sobre la base de las razones siguientes[14]:

— *Similaridad como coincidencia en todas las propiedades*: esta definición de similaridad no sirve, puesto que no hay ningún par de cosas que tengan todas sus propiedades en común, por lo que –en caso de aceptarla– ningún par de objetos serían similares.

— *Similaridad como coincidencia en alguna propiedad*: esta definición tampoco sirve, pues todo par de objetos coinciden en alguna propiedad (es decir, la similaridad es una relación universal), por lo que cualquier par de objetos serían similares entre sí y, por consiguiente, la noción de similaridad sería vacía[15,16].

— *Similaridad como relación triádica basada en propiedades*: la formulación triádica de similaridad de Quine –o, la más general definición *tetrádica* considerada por Goodman, del tipo "*a* y *b* son más similares que *c* y *d*"– tampoco serviría si esa mayor similaridad se explica en términos de propiedades. En este caso, la razón esgrimida por Goodman es que, dado que para un universo de *n* elementos cada par de ellos comparten 2^{n-2} propiedades, cuando el universo es infinito todo par de elementos de dicho universo compartirían el mismo –e infinito– número de propiedades[17].

— *Similaridad como coincidencia en propiedades relevantes* (o, similaridad en términos de la *importancia general de las propiedades compartidas*): Goodman considera que esta definición tampoco es adecuada, pues la relevancia –o importancia– de las distintas propiedades depende del contexto y los intereses del sujeto. No obstante, como veremos en la sección 3.3.4, los defensores de los enfoques geométricos han argumentado justamente en el sentido contrario, sosteniendo que la similaridad sí

[14] Goodman también rechaza el posible empleo de una noción intensional de propiedad. No obstante, dado que no proporciona argumentos al respecto, he preferido no incluir dicha crítica en esta enumeración.

[15] Estas dos primeras dificultades –asociadas a la posibilidad de definir la similaridad, bien como coincidencia en todas las propiedades, bien como coincidencia en alguna propiedad– constituyen una crítica directa a las nociones de similaridad de Leibniz y Carnap.

[16] Para una discusión de las dos asunciones subyacentes a esta crítica y la siguiente (a saber, [i] que toda propiedad puede caracterizarse mediante un conjunto de objetos, y [ii] que todo conjunto de objetos determina una propiedad) véase Decock y Douven (2011, pp. 68-69).

[17] En este punto Goodman aplica el *teorema del patito feo* –en inglés, *ugly duckling theorem*– de Watanabe (1969), por medio del cual éste mostraba que un patito feo y un cisne son tan similares entre sí como lo son dos cisnes, y cuya conclusión era que los juicios de similaridad siempre conllevan asunciones con respecto a las propiedades relevantes. Para un repaso de esta relatividad pragmática de la relación de similaridad, junto de posibles respuestas a la misma véase Niiniluoto (1987, pp. 35-38).

puede ser sensible al contexto e intereses del sujeto, en la medida en que las propiedades relevantes puedan serlo[18].

3.3. Modelos geométricos

En el último siglo el modelo geométrico ha sido una de las aproximaciones a la noción de similaridad que más influencia ha tenido –cuando no el enfoque dominante– (Carnap 1928; Coombs 1954; Shepard 1957, 1958; Torgerson 1958; Schreider 1975), estando implícito tanto en el trabajo de Quine (1969) como en las críticas de Goodman (1951, 1972). El propósito de los modelos geométricos es representar la semejanza entre objetos por medio de proximidades espaciales. En cierto modo los modelos geométricos contemporáneos actualizan propuestas previas que –en mayor o menor medida– también caracterizaban los estímulos mediante puntos localizados en un espacio (Shepard 1980). Así, por ejemplo, Newton (1704, I ii 6) representaba los colores del espectro visible sobre un círculo; Drobisch (1855 II §24) ubicaba los tonos puros –en acústica– sobre una hélice; Helmholtz (1867) y Schrödinger (1920) representaban los colores en una variedad riemanniana; y Henning (1916, pp. 94 y 506) localizaba olores y sabores dentro de un prisma y un tetraedro, respectivamente.

3.3.1 Distancia y axiomas métricos

Conforme se ha indicado, los *modelos geométricos* (también llamados *modelos espaciales*) representan las relaciones de similaridad mediante distancias en un espacio métrico. La definición matemática de espacio métrico es la de un conjunto de puntos que tiene asociada una función distancia (también llamada *métrica*), de modo tal que la distancia entre cualquier par de puntos a y b de ese conjunto se encuentra determinada por esa función. Formalmente, un espacio métrico queda definido por un par $\langle M,d\rangle$, en donde M es el conjunto de puntos y d es la métrica –o función distancia[19]– sobre M (esto es, d: $M\times M\rightarrow\mathbb{R}$, siendo $\mathbb{R}$ el conjunto de los números reales[20]). Finalmente, para que un espacio sea considerado métrico su función distancia ha de satisfacer las tres condiciones si-

[18] Los partidarios de enfoques alternativos al modelo geométrico también han sostenido que sus propuestas de similaridad pueden ser sensibles al contexto. Así, por ejemplo, desde el *modelo de rasgos* se ha argumentado que las propiedades relevantes pueden determinarse tanto en base a su tipicalidad en los objetos / conceptos considerados, como en función de su diagnosticidad (Tversky 1977). Por su parte, desde el ámbito de los *modelos de alineamiento* y *transformacionales* se ha apelado, para explicar la determinación de las propiedades y relaciones relevantes, tanto a su sistematicidad –o capacidad para formar parte de sistemas de relaciones más amplios– (Gentner 1983), como a su capacidad para producir analogías exitosas (Medin *et al.*, 1993).

[19] La función distancia entre dos puntos a y b –denotada como $d(a,b)$– también puede ser interpretada como la disimilitud existente entre esos puntos (o, alternativamente, como la disimilitud entre los objetos por ellos representados).

[20] En realidad la función distancia es una aplicación del producto cartesiano $M\times M$ en $\mathbb{R}_0^+$ (esto es, en el conjunto de los números reales positivos, incluyendo el cero), en la medida en que la condición de *positividad* –a saber, que toda distancia es mayor o igual que cero– se demuestra trivialmente a partir de las condiciones de minimalidad, simetría y desigualdad triangular.

guientes (Blumenthal 1953, p. 15) –también conocidas como *axiomas métricos* o *axiomas de distancia*[21]–:

$$d(a,b)=0 \text{ syss } a=b \ (\textit{minimalidad})$$
$$d(a,b)=d(b,a) \ (\textit{simetría})$$
$$d(a,b)+d(b,c)\geq d(a,c) \ (\textit{desigualdad triangular})$$

En consecuencia, minimalidad, simetría y desigualdad triangular son las tres asunciones básicas en un modelo geométrico estándar. El requisito de *minimalidad* exige que todo objeto sea mínimamente disimilar –o, alternativamente, máximamente similar– a sí mismo, y que lo sean todos ellos por igual[22]. La condición de *simetría* indica, por su parte, que el orden de los objetos considerados no afecta a la (di)similaridad existente entre ellos (esto es, que el objeto *a* es tan similar a *b* como el objeto *b* lo es a *a*). En último lugar, la

[21] Junto con los axiomas métricos, los modelos geométricos también aceptan la aditividad de segmentos, además de las asunciones dimensionales –a saber, dominancia, consistencia y transitividad–. Con respecto a la *aditividad de segmentos*, esta condición exige que las distancias a lo largo de un segmento sean aditivas, lo que equivale a que dado un segmento con extremos *a* y *c*, y siendo *b* un punto interior de dicho segmento, entonces $d(a,b)+d(b,c)=d(a,c)$ (Beals *et al.* 1968).

En cuanto a las *asunciones dimensionales*, su carácter no-métrico permite definirlas no solo en términos de una función distancia métrica *d*, sino también en términos de una medida ordinal de distancia δ. Dicho esto, las asunciones dimensionales –que permiten calificar de *monótona* a una estructura de proximidad, por lo que también son llamados *axiomas de monotonicidad*– pueden definirse como sigue (Tversky y Gati 1982, p. 124):

- *Dominancia*: una distancia bidimensional excede sus componentes unidimensionales, esto es:

$$\delta(x_1y_1,x_2y_2)>\delta(x_1y_1,x_1y_2) \text{ y } \delta(x_1y_1,x_2y_2)>\delta(x_1y_2,x_2y_2)$$

- *Consistencia*: el orden de los intervalos en un atributo no depende de los valores del otro atributo:

$$\delta(x_1y_1,x_2y_1)>\delta(x_3y_1,x_4y_1) \text{ syss } \delta(x_1y_2,x_2y_2)>\delta(x_3y_2,x_4y_2)$$
$$\delta(x_1y_1,x_1y_2)>\delta(x_1y_3,x_1y_4) \text{ syss } \delta(x_2y_1,x_2y_2)>\delta(x_2y_3,x_2y_4)$$

- *Transitividad*: la relación “estar entre” es transitiva (o no-circular). Si decimos que, para tres puntos alineados *a*, *b* y *c*, el punto *b* está entre *a* y *c* (relación denotada como $a|b|c$) cuando $\delta(a,c)>\delta(a,b)$ y $\delta(a,c)>\delta(b,c)$, entonces que la relación “estar entre” sea transitiva equivale a:

$$a|b|c \text{ y } b|c|d \rightarrow a|b|d \text{ y } a|c|d$$

En el caso particular de la métrica de Minkowski, las asunciones dimensionales son condiciones necesarias y suficientes para las que esa métrica cumpla las propiedades de *subtractividad intradimensional* (esto es, que la contribución de cada componente sea el valor absoluto de sus diferencias) y *aditividad interdimensional* (es decir, que la distancia sea una función de la suma de sus contribuciones componentes) (Tversky y Krantz 1970). De hecho, para el caso de un espacio *n*-dimensional euclidiano (que no con métrica euclidiana), las únicas métricas con segmentos aditivos que satisfacen la subtractividad intradimensional son las métricas generales de Minkowski (*ib.*, p. 589). Finalmente, para una explicación de las relaciones jerárquicas existentes entre todas estas asunciones y propiedades véase Borg y Groenen (1997, pp. 293-294).

[22] En este punto conviene indicar que la condición de minimalidad indicada es ligeramente más fuerte que la exigida por Tversky (1977, p. 328), y referida por él con el mismo nombre. De hecho, la condición de minimalidad de Tversky, a saber, $d(a,a)=0$, no garantiza que el espacio/distancia sea métrico –en contra de lo sostenido por Tversky–, sino tan solo pseudo-métrico (Kelley 1955, pp. 118-119).

desigualdad triangular exige que la disimilitud –o distancia– entre dos objetos *a* y *c* no pueda ser mayor que la disimilitud entre *a* y un tercer objeto *b* más la disimilitud entre *b* y *c*. En términos geométricos esto equivale a decir que el camino más corto entre dos objetos es el segmento de recta que conecta los puntos que representan tales objetos.

Sobre esta base, los modelos geométricos representan las entidades como puntos de un espacio métrico organizado en dimensiones, construido a partir de juicios de similaridad (o disimilitud), matrices de confusión, coeficientes de correlación, etc. El resultado de esta construcción es un espacio representacional en donde los puntos representan objetos, y la similaridad entre dos objetos –*a* y *b*, por ejemplo– es inversamente proporcional a la distancia *d* existente entre ellos. Así, Shepard describe[23] la similaridad entre dos objetos –o estímulos– *a* y *b* en términos de una función exponencial negativa de la distancia existente entre ellos[24]:

$$s(a,b) = \exp[-k \cdot d(a,b)]$$

En donde *k* es un parámetro que determina cómo de deprisa disminuye la similaridad entre dos objetos con la distancia existente entre ellos.

Finalmente, la distancia entre dos objetos *a* y *b* es normalmente computada como una distancia de Minkowski[25] la cual, en un espacio *n*-dimensional donde $x_i^{[u]}$ representara la posición del objeto *u* en la *i*-ésima dimensión[26], estaría dada por la expresión siguiente:

$$d(a,b) = \left(\sum_{i=1}^{n} \left| x_i^{[a]} - x_i^{[b]} \right|^p \right)^{1/p}$$

La métrica *p* de Minkowski –también referida como *métrica de potencias* o *métrica* L^p (Tversky y Krantz 1970)– define una familia de funciones (una para cada valor del parámetro $p \geq 1$) que expresan la distancia entre dos objetos por medio de las potencias de las diferencias entre sus componentes. Con ello, el valor del parámetro *p* determina el tipo de métrica y distancia: si p=1 estaríamos ante una métrica *city-block* (o Manhattan); si p=2 la métrica sería *euclidiana*.

[23] Sobre la base de un conjunto de estudios empíricos que le condujeron a proponer su *ley universal de generalización* para la ciencia cognitiva (Shepard 1980, 1987).

[24] El empleo de una función exponencial negativa hace que la variación de la similaridad –con respecto a la distancia– aumente conforme las distancias disminuyen.

[25] Para una presentación de las métricas de Minkowski, junto con muchas de sus propiedades formales véase Busemann (1955, pp. 94-104).

[26] Churchland, en su respuesta a las críticas de Fodor y Lepore (1992) en contra de la posibilidad de que un modelo neuronal pueda dar cuenta de nociones tales como *identidad* o *similaridad* conceptual (dada la gran diversidad estructural y funcional que –argumentan– puede esperarse encontrar en las redes neuronales constituyentes del cerebro de los distintos sujetos), plantea una interpretación "geométrica" análoga para los modelos de cognición basados en redes neuronales. Bajo dicha interpretación, los patrones de activación de una red neuronal pueden ser concebidos como puntos en un espacio (privado) *n*-dimensional, cuyas dimensiones (privadas) fuesen los niveles de activación de cada una de las *n* neuronas intervinientes (Churchland 1998, p. 6).

3.3.2 Escalamiento multidimensional

Históricamente, ha sido el escalamiento multidimensional (EMD) –en inglés, *multidimensional scaling* (MDS)– la aproximación preferida para caracterizar los modelos geométricos[27] (Torgerson 1952, 1965; Shepard 1962a, 1962b, 1980). El propósito del EMD es, sobre la base de las similaridades –o disimilitudes– observadas entre cada par de elementos pertenecientes a un cierto conjunto de objetos, encontrar una representación de dichos objetos mediante unas pocas dimensiones, de modo tal que las distancias entre cada par de objetos en el nuevo espacio encajen lo más posible con las distancias existentes entre ellos en el espacio original.

Dado un conjunto O de n objetos, la entrada del EMD será una matriz n-por-n de proximidades –o distancias– que describa cómo de cerca está cada objeto de O del resto de objetos pertenecientes a O. Dicha matriz de entrada representará a los n objetos por medio de $n(n-1)/2$ valores –o distancias–. Por su parte, la salida del EMD es una matriz n-por-m que constituye una representación geométrica de cada uno de los n objetos en un espacio m-dimensional, en este caso mediante nm valores. Sobre esta base, el EMD constituye un método para la compresión de información –al representar los datos de entrada por medio de un conjunto menor de dimensiones (dado que $m<n$)–. Además, la reducción dimensional revela los factores subyacentes a los datos de entrada, y facilita el poder dar –a dichos factores constituyentes del espacio representacional reducido– una interpretación psicológica[28].

Por ejemplo, si dispusiéramos de los siguientes valores de similaridad[29] entre China, Japón y Corea del Norte:

— Similaridad (China, Japón) = 3
— Similaridad (China, Corea del Norte) = 7
— Similaridad (Japón, Corea del Norte) = 1

El EMD intentaría construir un espacio tal que, cuando se ubiquen estos tres países en él, aquellos países juzgados más similares estuvieran más próximos entre sí que aquellos otros considerados menos similares. Obviamente, conforme aumente el número de dimensiones empleadas, aumentará la bondad de ajuste del modelo, pudiéndose alcanzar un ajuste perfecto si la dimensionalidad del modelo resultante es lo suficientemente alta. Para ello el modelo debería agotar el número de grados de libertad –o, parámetros que

[27] En todo caso, los algoritmos de agrupamiento –también llamados de *clustering*– son también una opción muy versátil a la hora de articular de este tipo de modelos.

[28] No obstante, uno de los problemas del EMD es que los juicios de similaridad (o datos de proximidades / distancias) entre los objetos no son un input disponible en el aprendizaje de los sujetos. Por esta razón, aún cuando el EMD puede ser útil a la hora de realizar un análisis –o reconstrucción– racional de cuál es la estructura del espacio conceptual de un sujeto, no constituiría un modelo válido de cómo opera realmente ese aspecto de la cognición (es decir, de cómo los conceptos se adquieren).

[29] Dados en una escala del 1 al 10, en donde 1 fuera baja similaridad y 10 fuera alta similaridad.

pueden variar de modo independiente– de los datos de entrada, por lo que el ajuste perfecto de un conjunto con n objetos precisaría de un espacio con n–1 dimensiones[30].

3.3.3 Puntos fuertes y débiles

La principal ventaja del modelo geométrico es que da perfecta cuenta de la noción de similaridad, tanto de los juicios de similaridad absolutos propios de la perspectiva clásica, como de la concepción comparativa de similaridad que Quine (1969) concebía en torno a relaciones de tipo triádico. En el primer caso –*similaridad absoluta*– dos objetos se considerarán similares si la distancia entre ellos cumple una cierta condición (por ejemplo, si no supera un determinado umbral). En el segundo caso –*similaridad comparativa*– dados tres objetos a, b, y c diríamos que a es más similar a b que a c si la distancia entre a y b es menor que la distancia entre a y c. En ambos casos el modelo geométrico proporciona una caracterización formal de las relaciones consideradas.

Por su parte, los métodos de EMD tienen tres aplicaciones, o ventajas, principales. Primero, permiten identificar las dimensiones subyacentes a un conjunto inicial de datos (en este caso, a los juicios de similaridad –o disimilitud– de entrada). En segundo lugar, permiten producir espacios conceptuales con una dimensionalidad mucho menor que la de las descripciones perceptuales iniciales, lo cual resulta cognitivamente eficiente en términos de codificación, memoria y procesamiento. En tercer lugar, el modelo resultante puede emplearse para caracterizar funciones cognitivas tales como categorizaciones, inferencias, memoria, etc.

No obstante, uno de los problemas del EMD es que, aún cuando puede ser un enfoque válido a la hora de intentar caracterizar la forma/organización de nuestros espacios conceptuales (sobre la base de un conjunto de juicios de similaridad/disimilitud obtenidos preguntando a los sujetos), resulta más cuestionable si se pretende emplear para explicar

[30] El motivo por el que un espacio con dimensionalidad n–1 permite una caracterización perfecta de las distancias entre n objetos es simple. Por un lado, el número de grados de libertad es igual al número de distancias existentes entre los objetos considerados. Así, el número de grados de libertad para un conjunto de n objetos será un número triangular dado por la fórmula $n(n–1)/2$. La razón es que si k objetos tienen n_k grados de libertad, la adición de un nuevo objeto añadirá k grados de libertad (sobre n_k), asociados a las distancias entre el nuevo objeto y los k objetos anteriores, por lo que la serie de números de grados de libertad no es otra que la serie de números triangulares.

Sobre esta base puede razonarse como sigue. Dados dos objetos a y b es posible situar al objeto a sobre una recta, y al objeto b a la distancia adecuada $d(a,b)$ sobre dicha recta. La adición de un tercer objeto c obliga, para su caracterización perfecta, a añadir una segunda dimensión. Los dos primeros objetos a y b se situarían del modo ya indicado sobre la recta (o hiperplano unidimensional), y el tercer objeto c se ubicaría –fuera de esa recta– en la intersección de las circunferencias centradas en a y b y con radios $d(a,c)$ y $d(b,c)$, respectivamente. La adición de un cuarto objeto d obliga, para una caracterización perfecta, a añadir una tercera dimensión. Los tres primeros objetos a, b y c se ubicarían sobre un plano (o hiperplano bidimensional) del modo antes indicado, y el cuarto objeto d se situaría –fuera de ese plano– en la intersección de las superficies esféricas centradas en a, b y c y con radios $d(a,d)$, $d(b,d)$ y $d(c,d)$, respectivamente. Como ya debería ser obvio, los k nuevos grados de libertad introducidos por cada nuevo objeto con respecto al conjunto de k objetos previos situados en un hiperplano (k–1)dimensional, obligan a introducir una nueva dimensión k en donde la localización del nuevo objeto quedará determinada por sus k distancias con respecto a los objetos anteriores.

cómo los sujetos adquieren los conceptos, o cuál es el origen de sus elementos constitutivos más básicos (esto es, cuál es el origen de las dimensiones que constituyen esos espacios conceptuales). En este caso la cuestión crucial es que la matriz de similaridades –o disimilitudes– es una entrada necesaria para que el EMD pueda producir un modelo, razón por la cual dicha matriz deberá estar disponible al inicio del proceso de análisis. Sin embargo, aún cuando esa matriz es un input no problemático en estudios psicológicos (en los que se pregunta a los sujetos sobre sus juicios de similaridad acerca de un cierto conjunto de objetos), tal no es el caso cuando se considera –por ejemplo– cómo un sujeto determina las dimensiones constitutivas de uno de sus espacios conceptuales pues, en este segundo caso, la asunción de que la matriz de similitudes es un input disponible para el sujeto en su proceso de aprendizaje resulta mucho más controvertida.

Y, ya de modo general, otra objeción que –en ocasiones– se hace a los modelos geométricos es que los juicios de similaridad a menudo dependen del contexto, razón por la cual las circunstancias podrían alterar las similaridades percibidas/evaluadas para un mismo conjunto de objetos (Goodman 1972, p. 445). En este caso, la crítica es doble. Por un lado, pone en cuestión que los modelos de tipo geométrico dispongan de parámetros que los hagan sensibles al contexto. Por el otro, y lo que es más grave, el carácter relativo de la similaridad –con respecto a un contexto o propiedades relevantes– la convierte en una noción vaga si no se concreta cuál es el contexto del discurso (esto es, con respecto a qué propiedad(es) son dos objetos similares). Pero, si a la afirmación de que dos cosas son similares le añadimos la especificación de la propiedad que comparten, entonces –en opinión de Goodman– la afirmación de su similaridad se convierte en superflua[31]. Frente a esta crítica, Goldstone y Son (2005) consideran que hay buenas razones para rechazar la conclusión de Goodman –de que la similaridad es bien vaga o innecesaria–, en la medida en que (i) no siempre podemos dar cuerpo a la cláusula "con respecto a la propiedad *i*" mediante una única propiedad[32], y en esos casos el empleo de la similaridad resulta natural y primitiva (Smith y Kemler 1978; Smith 1989); y (ii) los adultos pueden tener impresiones generales de similaridad sin atender a propiedades específicas (Ward 1983; Smith y Kemler 1984).

3.3.4 Violación de los axiomas: contraejemplos y réplicas

En todo caso, posiblemente el mayor problema del modelo geométrico estándar son los tres axiomas métricos en los que se apoya pues, desde el trabajo de Tversky (1977), se ha sostenido que existen evidencias empíricas en contra de todos ellos:

— *Minimalidad*: en cuanto a la condición de minimalidad, su problema es que implica una misma similaridad reflexiva para todos los objetos. O, dicho de otro mo-

[31] A esto es precisamente a lo que luego apuntará Medin cuando indique que la similaridad resulta útil únicamente en la medida en que pueda concretar (a) los principios que determinan lo que es una propiedad relevante, y (b) los principios que determinan la importancia de cada propiedad particular (Medin 1989, p. 1474). Además, considerado esto, el trabajo explicativo recaería, no tanto en la noción de similaridad, como sí en los principios que determinan las propiedades relevantes y su importancia, por lo que la idea de similaridad sería más bien una variable dependiente que independiente.

[32] Este fenómeno resulta especialmente evidente en el caso de los niños.

do, implica que la similaridad de un objeto *a* con respecto a sí mismo deba ser igual a la similaridad de cualquier otro objeto *b* con respecto a sí mismo. En este caso el problema es que existen estudios empíricos –basados en el tiempo empleado para reconocer dos objetos como similares– que muestran que los sujetos identifican más rápidamente la similaridad entre dos ejemplares iguales de la letra **S**, que entre dos ejemplares iguales de la letra **W** (Podgorny y Garner 1979, p. 41). Y, aún peor, que la similaridad entre un ejemplar de la letra **C** y un ejemplar de la letra **O** es mayor que la similaridad entre dos ejemplares iguales de la letra **W** (*ib*., p. 47); e incluso que la letra **M** es más reconocida como una **H** (p=39.1%) que como una **M** (p=11%) (Gilmore *et al.* 1979, p. 427).

— *Simetría*: con respecto al axioma de simetría Tversky ya demostraba que los juicios de los sujetos muchas veces no son simétricos, tal y como sucedía en el caso de China y Corea del Norte en donde la mayoría de los sujetos asentían a la afirmación “Corea del Norte es similar a China”, mientras que muy pocos de ellos asentían a la afirmación simétrica “China es similar a Corea del Norte” (Tversky 1977, p. 334; Tversky y Gati 1978, pp. 84-85). En este caso Tversky y Gati argumentaban que las asimetrías eran debidas a la distinta importancia que tienen los países en ambas oraciones[33]. Otras violaciones semejantes de la condición de simetría también han sido observadas en contextos de tipo social, en los que –por ejemplo– los sujetos juzgan que sus amigos son más similares a ellos mismos que lo que ellos lo son con respecto a esos mismos amigos (Holyoak y Gordon 1983).

— *Desigualdad triangular*: el problema de este tercer axioma es que Tversky y Gati (1982) realizaron una serie de estudios empíricos en los que mostraban que esta condición no era compatible con la condición de segmentos aditivos. La condición de aditividad de segmentos exige que las distancias a lo largo de un segmento sean aditivas –esto es, dado un par de puntos *a* y *c*, y siendo *b* un punto perteneciente al segmento *ac*, entonces $d(a,b) + d(b,c) = d(a,c)$– (Beals *et al.* 1968)[34]. Ahora bien, la combinación de la desigualdad triangular con la aditividad de segmentos para el caso de un triángulo rectángulo con vértices *a*, *e* y *c*, cuya hipotenusa (segmento *ac*) contuviese un cuarto punto *b*, permite obtener la desigualdad $d(a,e) + d(e,c) \geq d(a,b) + d(b,c)$, o *desigualdad de esquina* (ver Fig. 3.1). Esta desigualdad se cumple si el camino de esquina excede el camino central, lo cual ocurrirá si $d(a,e) \geq d(a,b)$ y $d(e,c) \geq d(b,c)$, o si $d(a,e) \geq d(b,c)$ y $d(e,c) \geq d(a,b)$; y no se cumple si el camino central es mayor que el camino de esquina, lo cual sucede si se satisface una condición igual a la anterior pero con desigualdades opuestas[35]. El problema es que la investigación

[33] Su tesis era que en estas afirmaciones el sujeto es el país menos prominente, mientras que el otro actúa como referente en la evaluación de la similaridad. Debido a ello el segundo país (esto es, el más prominente) establecería el contexto en el que la similaridad se evalúa y, siendo contextos distintos, no es extraño que sus similaridades asociadas también difieran.

[34] Véase nota al pie 21 de este capítulo.

[35] Estas dos condiciones no agotan los casos en los que la desigualdad de esquina se cumple –o no se cumple–, aunque sí aquellos en los que su cumplimiento –o incumplimiento– puede comprobarse con disimilaridades puramente ordinales.

de Tversky y Gati (1982, pp. 131-132) mostraba violaciones sistemáticas de esta desigualdad (en donde el camino central era significativamente mayor que el camino de esquina) lo que no solo refutaba la desigualdad de esquina, sino que –con ello– probaba que la desigualdad triangular era incompatible con la aditividad de segmentos.

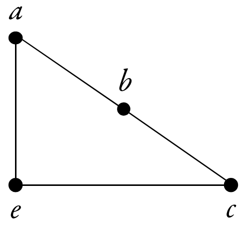

Fig. 3.1. Triángulo rectángulo y esquema de puntos considerados en la desigualdad de esquina.

Por su parte, los defensores de los modelos geométricos han respondido a las anteriores violaciones de los axiomas métricos apelando a la noción de contexto. Este tipo de respuestas parten de la asunción de que los juicios de similaridad tienen lugar siempre en el seno de un contexto, sin el cual no podrían evaluarse. O, en otras palabras, la similaridad entre objetos únicamente podría determinarse con respecto a ciertos aspectos (o propiedades), y esos aspectos relevantes pueden variar de contexto en contexto[36]. Ahora bien, conforme veremos a continuación no existe un único modo de caracterizar la influencia del contexto sobre la similaridad, sino múltiples propuestas alternativas al respecto.

Comenzando con la violación de los *axiomas de minimalidad* y *simetría*[37], las principales respuestas a ellas recurren, bien a la noción de densidad (Krumhansl), bien a la de sesgo (Holman y Nosofsky). Así, Krumhansl (1978) considera que la disimilaridad entre dos estímulos es una función creciente, tanto de la distancia entre los puntos que los representan, como de la densidad de estímulos en torno a cada uno de esos dos puntos, los cuales en este caso actuarían como contexto. Aquí "estímulos" debe entenderse en el sentido de puntos representando objetos del mundo a los que el sujeto haya estado expuesto, por lo que "densidad de estímulos" refiere a la concentración de puntos en una cierta región del espacio representacional. Sobre esta base, la función distancia modificada $d^*(a,b)$ considerada por Krumhansl adoptaría la forma siguiente:

$$d^*(a,b) = d(a,b) + \alpha\delta(a) + \beta\delta(b)$$

En donde $\delta(x)$ es una medida de la densidad espacial en el entorno del punto x, y los parámetros α y β serían los pesos dados a las densidades $\delta(a)$ y $\delta(b)$, respectivamente. En este modelo dos pares de puntos (a,b) y (c,d) situados a igual distancia –esto es, tales que

[36] Esta respuesta –a saber, una concepción contexto-dependiente de la similaridad– permite a los modelos geométricos, no solo dar cuenta de las violaciones de los axiomas de distancia, sino también satisfacer la dependencia del contexto –o circunstancias– puesta de manifiesto por Goodman (1972, p. 445).

[37] Para un repaso completo de las aproximaciones y modelos de similaridades –o proximidades– asimétricas véase Zielman y Heiser (1996).

$d(a,b) = d(c,d)$– pero localizados en regiones con distintas densidades de estímulos (por ejemplo, *a* y *b* en una región más densa, y *c* y *d* en una región menos densa) tendrían similaridades distintas, y tales que $s(a,b) < s(c,d)$. La ventaja de este enfoque es que da cuenta del hecho de que los sujetos hagan discriminaciones más finas de similaridad en regiones más densas que en aquellas otras menos densas. Así, si en el ejemplo antes presentado de violación del axioma de minimalidad, la letra W estuviera situada en una región más densa que la letra S, la similaridad entre dos ejemplares iguales de la letra W sería menor que la similaridad entre dos ejemplares iguales de la letra S, lo cual explicaría por qué estos últimos –letras S– son identificados más rápidamente (como similares) por los sujetos que los dos primeros –letras W–[38]. Por tanto, el hecho de que en el modelo de Krumhansl la auto-similaridad (o similaridad de un objeto consigo mismo) sea mayor para objetos representados por puntos localizados en regiones poco densas, que para objetos representados por puntos en regiones más densas, explicaría la violación del *axioma de minimalidad*.

Además, debido al carácter direccional de los juicios de (di)similaridad, en donde uno de los objetos actúa como sujeto y el otro como referente (Tversky 1977, p. 328) (por ejemplo, en el juicio "*a* es similar a *b*", el objeto *a* sería el sujeto, y el objeto *b* sería el referente), la contribución a la disimilaridad de la densidad en torno a cada objeto podría estar diferentemente ponderada para el sujeto y el referente, lo que daría lugar a la existencia de similaridades asimétricas y, por consiguiente, explicaría la violación del *axioma de simetría*. Así, por ejemplo, en el caso de la similaridad entre China y Corea del Norte, si sus densidades espaciales fuesen δ(China)=4 y δ(Corea del Norte)=1, y sujeto y referente recibiesen pesos distintos $\alpha=3$ y $\beta=1$, el modelo de Krumhansl explicaría por qué los sujetos asienten más frecuentemente al juicio "Corea del Norte es similar a China", que al juicio "China es similar a Corea del Norte".

Por su parte, las respuestas de Holman (1979) y Nosofsky[39] (1991) a la violación del axioma de simetría se basan en una función de sesgo o, por comparación con el modelo de Krumhansl, consideran que la similaridad entre dos estímulos es función, tanto de la distancia entre los puntos que la representan, como del sesgo de cada uno de esos estímulos[40]. En este caso los dos estímulos considerados serían el factor que actúa como contexto, el cual estaría asociado al hecho de que un cierto estímulo pueda ser más prominente en percepción o memoria, más fácilmente atendido, más rápidamente codificado[41], etc. Tal y como ocurría en el modelo de Krumhansl, el carácter direccional de los juicios de similaridad permite concebir que el sesgo de cada estímulo pueda depender de su rol en el juicio de similaridad, lo que permitiría la existencia de similaridades asimétricas y, con

[38] Razonamientos análogos explican el resto de violaciones del axioma de minimalidad antes mencionadas.

[39] Algo más recientemente, Johannesson (2000) ha propuesto un modelo de prominencias relativas, que no es más que una versión logarítmica del modelo basado en sesgos de Nosofsky.

[40] Este "sesgo del estímulo" en ocasiones también ha sido llamado "peso del estímulo" (Shepard 1957).

[41] O, en otras palabras, el sesgo de un estímulo sería una medida de cómo de bueno es ese estímulo, en donde la *bondad del estímulo* es indicativa de la eficiencia del sujeto al procesarlo (Garner y Clement 1963; Garner 1970).

ello, la violación del *axioma de simetría*. Por otro lado, el sesgo de estímulo también podría explicar las violaciones del *axioma de minimalidad*, en la medida en que los tiempos necesarios para el reconocimiento de similaridades podrían variar en función de los sesgos asociados a los estímulos considerados. Una propuesta similar a los modelos basados en sesgos es la de Takane y Sergent (1983), quienes definen la disimilaridad entre dos estímulos como una función de la distancia que los separa y la complejidad de cada uno de ellos, la cual también permitiría explicar la violación del *axioma de minimalidad*[42].

En cuanto a las violaciones de la *desigualdad triangular*, en este caso se ha argumentado que el contexto podría no activar todas las dimensiones asociadas al conjunto de conceptos considerado o, alternativamente, que tales dimensiones pueden tener pesos distintos en función del contexto (Krumhansl 1978; Holman 1979; Nosofsky 1991). En ambos casos, la función de similaridad resultante podría no cumplir la desigualdad triangular. Finalmente, otra posible explicación de la violación de la desigualdad triangular es la propuesta por Nosofsky (1992b) –aunque ya tácitamente presente en Tversky y Gati (1982, p. 150)–, según la cual el valor $p < 1$ obtenido por Tversky y Gati para el modelo de potencias de Minkowski (y en virtud del cual concluían el no cumplimiento de la desigualdad triangular) podría explicarse –o interpretarse– como que los sujetos proporcionan sistemáticamente más atención (es decir, ponderan más) a aquellas dimensiones en las que los estímulos son más similares[43].

3.4. Modelos de rasgos

Todas las anteriores aparentes deficiencias del modelo geométrico condujeron a Tversky (1977) a proponer una caracterización de la similaridad en términos de un proceso de emparejamiento –en inglés, *matching*– de rasgos. El *modelo de rasgos* de Tversky (también llamado, *modelo de contraste*) es una aproximación a la similaridad basada en la teoría de conjuntos, en donde las entidades (objetos y conceptos) se representan por medio de conjuntos de rasgos –o atributos– constitutivos[44].

3.4.1 Axiomas y teorema de representación

Tversky presentaba su propuesta en los términos siguientes. Sea un dominio de objetos o estímulos $\Delta = \{a, b, c, \ldots\}$; sean A, B, C, etc., los conjuntos de rasgos asociados con los objetos a, b, c, etc., respectivamente; y sea $s(a,b)$ la medida de la similaridad entre los objetos a

[42] Otra explicación alternativa de cómo el contexto puede dar cuenta del no cumplimiento del *axioma de minimalidad* puede encontrarse en Ashby y Perrin (1988, pp. 133-134).

[43] Esta interpretación sería consistente con la sospecha de Sjöberg y Thorslund (1979) de que los sujetos buscan activamente modos en que los estímulos de entrada generen clases de objetos altamente similares. En el caso de la explicación de Nosofsky, la mayor ponderación de las dimensiones en las que los estímulos son más similares contribuiría precisamente a ese propósito.

[44] Por ejemplo, el concepto PLÁTANO podría tener como atributos constitutivos los pertenecientes al conjunto {*amarillo*, *falcado*, *dulce*, *tamaño-medio*}; mientras que un cierto plátano concreto, en cambio, podría estar representado por el conjunto de rasgos {*verde*, *falcado*, *amargo*, *tamaño-medio*}.

y *b*. Sobre esta base, el modelo de rasgos se articula en torno los siguientes tres axiomas principales:

(1) *Emparejamiento*: la similaridad entre los objetos *a* y *b* es función (i) de los rasgos comunes entre *a* y *b* ($A \cap B$); y (ii) de sus rasgos distintivos, a saber, de los rasgos de *a* que no son rasgos de *b* (o diferencia entre los rasgos de *a* y *b*, esto es, $A-B$), y de los rasgos de *b* que no son rasgos de *a* (o diferencia entre los rasgos de *b* y *a*, esto es, $B-A$). Esto suele expresarse del modo siguiente:

$$s(a,b)=f(A \cap B, A-B, B-A)$$

Los componentes –o factores– de la función *f* (a saber, $A \cap B$, $A-B$ y $B-A$) no son más que subconjuntos del conjunto de todos los rasgos asociados a los objetos del dominio Δ, denotados como *X*, *Y* y *Z* (donde $X=A \cap B$, $Y=A-B$ y $Z=B-A$).

(2) *Monotonicidad*: según la cual la función de similaridad aumenta con la inclusión de rasgos comunes y con la supresión de rasgos distintivos. Esta condición puede formularse de la manera siguiente[45]:

$$s(a,b) \geq s(a,c) \quad \text{si} \quad A \cap C \subseteq A \cap B \;\; \text{y} \;\; A-B \subseteq A-C \;\; \text{y} \;\; B-A \subseteq C-A$$

Cualquier función *f* que cumpla los axiomas de emparejamiento y monotonicidad es llamada por Tversky (*ib.*, p 330) *función de emparejamiento*, pues permite determinar el grado en que dos objetos –concebidos en términos de conjuntos de rasgos– encajan entre sí.

(3) *Independencia*: el axioma de independencia sostiene que el orden del efecto conjunto de cualesquiera dos componentes de la función *f* es independiente del valor fijo del tercer factor. Esto equivale a que, si suponemos que los pares (a,b) y (c,d), y los pares (a',b') y (c',d') comparten las mismas dos componentes –sean éstas cuales sean–; y que los pares (a,b) y (a',b'), y los pares (c,d) y (c',d') comparten la restante tercera componente; entonces:

$$s(a,b) \geq s(a',b') \quad \text{syss} \quad s(c,d) \geq s(c',d')$$

A partir de estos axiomas[46], Tversky prueba su *teorema de representación*, según el cual existe una escala de similaridad *S* y una escala no-negativa *F* tales que, para cualesquiera cuatro objetos *a*, *b*, *c* y *d* del dominio Δ, se cumplen las tres condiciones siguientes:

[45] En esta expresión la desigualdad no-estricta ≥ puede substituirse por la desigualdad estricta > si al menos una de las tres relaciones de inclusión ⊆ es una relación de inclusión propia ⊂.

[46] Además de los tres axiomas principales (de *emparejamiento*, *monotonicidad* e *independencia*), el modelo de rasgos de Tversky asume otros dos axiomas de tipo instrumental. El cuarto axioma –o *solubilidad*– precisa que el espacio de rasgos sea lo suficientemente rico como para que ciertas ecuaciones de similaridad puedan resolverse. Por su parte, el quinto axioma –o *invariancia*– garantiza que la equivalencia de intervalos se preserva a través de las distintas componentes (o factores). Para una presentación detallada de la formulación axiomática del modelo de rasgos, junto con la prueba de su teorema principal, véase Tversky (1977, pp. 350-351).

$$S(a,b) \geq S(c,d) \quad \text{syss} \quad s(a,b) \geq s(c,d)$$
$$S(a,b) = \theta F(A \cap B) - \alpha F(A-B) - \beta F(B-A), \text{ con } \theta, \alpha, \beta \geq 0$$
F y S son escalas de intervalo

Lo que el teorema de representación dice es que existe una escala de similaridad S que preserva el ordenamiento de similaridad observado, y que dicho orden se puede expresar como una combinación lineal –o contraste (de ahí que este enfoque también se conozca como *modelo de contraste*[47])– de sus rasgos comunes y distintivos. En la expresión anterior los parámetros θ, α y β son los pesos de los factores comunes y distintivos, en función de la importancia dada a cada uno de ellos por el sujeto en los diferentes contextos.

Por tanto, el modelo de rasgos opera mediante un proceso de encaje –o emparejamiento– de rasgos basado en la diferente ponderación de los rasgos comunes ($A \cap B$) y los dos tipos de rasgos distintivos ($A-B$ y $B-A$), siendo habitual dar mayor peso a los rasgos comunes –parámetro θ–, que a los distintivos –parámetros α y β–.

3.4.2 Ventajas frente al modelo geométrico

Originalmente los modelos de rasgos presentaban[48] dos ventajas principales frente a los modelos geométricos, a saber, (i) eran capaces de explicar la dependencia contextual de los juicios de similaridad, y (ii) predecían la existencia de similaridades asimétricas. En cuanto a la dependencia del contexto, los modelos de rasgos representan los objetos en un dominio mediante un subconjunto de todas las propiedades que cada objeto tiene –esto es, mediante sus propiedades relevantes según los intereses y propósitos del sujeto (los cuales variarán de contexto en contexto)–. Además, incluso dado un cierto conjunto de propiedades relevantes, la importancia de cada una de ellas podría variar de un contexto a otro. Finalmente, la propia selección del dominio podría ser sensible al contexto; así, por ejemplo, si todos los objetos considerados compartiesen un mismo rasgo, ese rasgo tendería a convertirse en neutral –Tversky lo llama *efecto de extensión*–y, por lo tanto, el domi-

[47] No obstante, el *modelo de contraste* no es el único modelo de rasgos posible, en la medida en que la función S puede adoptar otras formas alternativas, distintas de la considerada por Tversky en su teorema de representación. Así, por ejemplo, si la función de similaridad estuviese normalizada (de manera que variase entre 0 y 1) conforme a la expresión siguiente, nos encontraríamos ante un modelo de rasgos diferente –al que Tversky (1977, p. 333) da el nombre de *modelo de proporción*–:

$$S(a,b) = F(A \cap B) \ / \ [F(A \cap B) + \alpha F(A-B) + \beta F(B-A)], \text{ con } \alpha, \beta \geq 0$$

Con respecto a esto, existen múltiples antecedentes similares al modelo de contraste que caracterizan la similaridad en base a sus componentes comunes y distintivas. Así, por ejemplo, Bush y Mosteller (1951) definen la similaridad como $F(A \cap B)/F(A)$; por su parte, Ekman y colaboradores la conciben como $F(A \cap B)/(F(A)+F(B))$ –bajo la asunción de estímulos unidimensionales en donde la magnitud de A es menor que la de B– (Eisler y Ekman 1959; Ekman *et al.* 1964); mientras que Sjöberg la define como $F(A \cap B)/F(A \cup B)$ (Sjöberg 1972; Sjöberg y Thorslund 1979). Obviamente, todas estas aproximaciones caracterizan la similaridad como un ratio entre los rasgos comunes y distintivos (y no como una combinación lineal suya, tal y como hace el modelo de contraste), por lo que serían variaciones del modelo de proposición, que no del modelo de contraste.

[48] Digo que “presentaban” pues, como se ha visto en la sección anterior, es posible introducir variaciones en el modelo geométrico original –o estándar– que permiten incluir la influencia del contexto.

nio no lo incluiría dentro de su conjunto de propiedades relevantes (Tversky 1977, p. 343; Medin *et al.* 1993).

El segundo y más importante punto fuerte de los modelos de rasgos es su capacidad para dar cuenta de la violación de la condición de simetría en los juicios de similaridad. Nada en el modelo de contraste exige que los parámetros α y β sean iguales, ni tampoco que las componentes $F(A\text{–}B)$ y $F(B\text{–}A)$ tengan que serlo, lo que explicaría tanto las violaciones del axioma de simetría, como el carácter direccional de los juicios de similaridad. Así, por ejemplo, en la comparación entre China –sujeto, u objeto *a*– y Corea del Norte –referente, u objeto *b*–, si China tuviese más rasgos relevantes que Corea del Norte, y el parámetro α fuese mayor que β, eso explicaría que los sujetos juzguen que Corea de Norte es más similar a China que lo que China lo es a Corea del Norte.

Finalmente, el modelo de contraste también explica el carácter no especular de los juicios de similaridad y diferencia. En este caso, y bajo la asunción de que el peso dado a la componente común fuese mayor en los juicios de similaridad que en los de diferencia, y que los pesos de las componentes distintivas fuesen mayores en los de diferencia, entonces un par de estímulos podría ser percibido al mismo tiempo como más similar y más diferente entre sí que otro par de estímulos dado (Tversky 1977, p. 340; Medin *et al.* 1993). Esto era justamente lo que sucedía para los países Alemania Occidental y Alemania Oriental, los cuales eran considerados por una mayoría de sujetos como más similares (67% de los sujetos) y más diferentes (70% de los sujetos) entre sí que los países Ceilán y Nepal. En este caso, la explicación dada por Tversky es que el número de rasgos, en común y distintivos, conocidos por los sujetos para el caso de Alemania Occidental y Oriental es mayor que el número de rasgos comunes y distintivos conocidos para el caso de Ceilán y Nepal.

3.4.3 Problemas y limitaciones

No obstante, el modelo de rasgos no se encuentra libre de problemas, y el primero de ellos está asociado a cómo este enfoque representa las propiedades de los objetos, a saber, mediante conjuntos de rasgos (frente a los modelos geométricos, que las representaban como coordenadas en un espacio métrico). En este caso la dificultad surge al considerar cómo representar las propiedades continuas de los objetos mediante atributos binarios (que son los que emplea Tversky). Y, aunque los defensores del modelo de rasgos han planteado propuestas con respecto a cómo podrían caracterizarse con rasgos binarios variables tanto nominales –mediante el recurso a variables tipo *dummy*– como ordinales/continuas –bajo la forma de cadenas o anidamientos de rasgos– (Tversky y Gati 1982), conforme indican Hahn y Chater (1997) dichas respuestas resultan artificiosas e introducen un gran número de rasgos adicionales e innecesarios.

Otro problema de los modelos de rasgos, también mencionado por Hahn y Chater, es que, dado que la similaridad entre dos objetos es una función creciente del número de rasgos considerado, eso conduce al resultado de que la auto-similaridad –es decir, la similaridad de algo consigo mismo– de un objeto podrá variar en función del número de rasgos elegido. Esto, cuando se considera la evaluación de la auto-similaridad de un mismo objeto en circunstancias distintas, resulta una conclusión bastante implausible.

Finalmente, el modelo de rasgos comparte con los modelos geométricos una misma limitación, asociada al hecho de que en ambos casos sus representaciones son relativamente poco estructuradas (puesto que son meras colecciones de rasgos o coordenadas, respectivamente), lo cual constituye un problema cuando se considera el hecho de que en muchos casos es necesario que tales propiedades estén organizadas conforme a una cierta estructura (Hahn y Chater 1997; Goldstone y Son 2005). Esto es, una criatura con *pico*, *ojos*, *alas* y *cola* –si éstos fuesen los atributos del concepto PÁJARO–, pero que los tuviera colocados en la posición equivocada, difícilmente sería llamada *pájaro*.

3.5. *Modelos de alineamiento*

Los modelos basados en alineamientos surgieron con objeto de superar las dificultades que presentaban el modelo geométrico y el de rasgos a la hora de dar cuenta del modo en que los atributos de los objetos están interrelacionados y organizados. La captura de este tipo de relaciones requería de representaciones estructuradas, que tuvieran en consideración cómo las distintas partes –o elementos– de los objetos se relacionan, corresponden o alinean entre sí. Esto dio lugar a la aparición, en primera instancia, de los *modelos de alineamiento* (Markman y Gentner 1993a, 1993b; Goldstone 1994a; Gentner y Markman 1994, 1997) y, posteriormente, de los *modelos transformacionales* (de los que me ocuparé en la sección siguiente).

3.5.1 *Noción de alineamiento*

Los modelos de alineamiento estructural, a diferencia de los modelos geométricos y de rasgos, no se limitan a evaluar el encaje (en un cierto objeto) de un conjunto de atributos dados, sino que también comprueban si esos atributos se encuentran organizados del modo adecuado. O, dicho de otro modo, comprueban que los rasgos alineados (de los objetos y/o conceptos considerados) encajan, en el sentido de que desempeñen papeles similares dentro de sus respectivas entidades[49]. En tal caso, el correcto encaje de rasgos alineados aumentaría el peso de tales rasgos en el cómputo de la similaridad existente entre los objetos considerados. Así, por ejemplo, un *camión* con una rueda *blanca*, y un *coche* con una puerta *blanca*, comparten ambos el atributo BLANCO; no obstante, dado que los elementos de esos objetos que comparten ese atributo –a saber, *rueda* y *puerta*– no se encuentran alineados, entonces dicho rasgo compartido (esto es, el encaje del atributo BLANCO) no contribuiría a aumentar la similaridad entre los objetos considerados o, alternativamente, no contribuiría tanto como si estuviera asociado a dos elementos alineados[50, 51].

[49] Los modelos de alineamiento se inspiran en –o derivan de– los enfoques de razonamiento analógico, y su funcionamiento basado en mapeos estructurales que maximizan las correspondencias relacionales entre los elementos de los objetos comparados (Gentner 1980, 1983; Gentner y Toupin 1986; Holyoak y Thagard 1989, 1995).

[50] Otro posible ejemplo sería el puesto por Markman y Gentner (1993b, p. 434), en relación con los dos pares de objetos (Sol, Júpiter) y (Júpiter, Io). En este caso, si a un sujeto se le pidiera que identificase en el segundo par, (Júpiter, Io), el elemento más similar a Júpiter en el primer par, la respuesta trivial –en base a la similaridad de sus atributos– es que ese elemento es Júpiter (pues no hay nada más similar a Júpiter que él mismo). No obstante, cuando se pone de manifiesto la relación presente en ambos pares –a saber, que la causa por la que un elemento del par orbita alrededor del otro es que este último tiene

Probablemente, los dos modelos más importantes para la identificación de alineamientos estructurales sean el motor de mapeo de estructuras –en inglés, *Structure Mapping Engine*, o modelo SME– desarrollado por Falkenhainer, Forbus y Gentner (1986, 1989), y el modelo de mapeo y activación interactivos –en inglés, *Similarity as Interactive Activation and Mapping*, o modelo SIAM– planteado por Goldstone (1994a)[52].

3.5.2 Modelo SME

En cuanto al primero, el *modelo SME* es una extensión directa de la teoría de mapeo de estructuras por analogía propuesta por Gentner (1980, 1983). El modelo SME almacena la información –o representa el conocimiento– con una notación similar a la del cálculo de predicados, y distingue tres tipos de constructos principales (Falkenhainer *et al.* 1989): (a) *entidades* –u objetos–, que serían individuos y constantes; (b) *predicados*, que podrían ser funciones, atributos o relaciones; y (c) *D-grupos* –o *descripciones de grupos*–, que serían representaciones jerárquicas de conjuntos de entidades –o de hechos sobre ellas– tomados como una unidad. Sobre esta base, el motor recibiría como entrada para la comparación dos D-grupos (referidos como fuente –o base– y objetivo), tras lo cual calcularía sus correspondencias y buscaría el mejor encaje estructuralmente consistente. Así, por ejemplo, para el caso de la analogía utilizada por George Bush en 1991, los grupos fuente y objetivo podrían representarse como sigue[53]:

FUENTE:
Führer-de (*Hitler, Alemania*)
ocupar (*Alemania, Austria*)
mal (*Hitler*)
causa [*mal*(*Hitler*), *ocupar*(*Alemania, Austria*)]
primer-ministro-de (*Churchill, Gran Bretaña*)

más masa que el primero–, eso abre la puerta a otra posible identificación, a saber, la de Júpiter (en el primer par) con Io (en el segundo).

[51] Con respecto a las evidencias en favor de los modelos basados en alineamientos, cabe enumerar las siguientes: (a) El estar expuesto a múltiples ejemplos de una misma estructura –o esquema– de alineamiento ayuda a los sujetos a identificar y aplicar esa estructura en el futuro (Gick y Holyoak 1983; Catrambone y Holyoak 1989). (b) La comparación explícita de ejemplos facilita la identificación de estructuras de alineamiento comunes, más que la mera presentación de los mismos ejemplos (Loewenstein y Gentner 2001, 2005). (c) Los sujetos realizan alineamientos de modo progresivo; así, por ejemplo, la tasa de éxito con que los sujetos identifican un alineamiento entre objetos disimilares aumenta si antes han realizado dicho mismo alineamiento para otros objetos más similares (Gentner, Loewenstein y Hung 2007).

[52] Otros modelos destacados para la determinación de alineamientos estructuralmente consistentes, aunque –en este caso– basados en mapeos de analogías, serían el modelo ACME de Holyoak y Thagard (1989), el proyecto Copycat de Hofstadter y Mitchell (1994), el modelo LISA de Hummel y Holyoak (1997) y el modelo IAM de Keane y colaboradores (1994). Para una más completa revisión de los modelos computacionales de analogías véase French (2002).

[53] Este ejemplo ha sido tomado de Holyoak (2005, p. 132).

causa [*ocupar*(*Alemania, Austria*), *contraatacar*(*Churchill, Hitler*)]

OBJETIVO:
presidente-de (*Hussein, Irak*)
invadir (*Irak, Kuwait*)
mal (*Hussein*)
causa [*mal*(*Hussein*), *invadir*(*Iraq, Kuwait*)]
presidente-de (*Bush, Estados Unidos*)

En el ejemplo anterior encontramos individuos (tales como *Hitler* o *Churchill*), atributos (o predicados unarios, tales como *mal*) y relaciones (de primer orden, como por ejemplo *presidente-de* u *ocupar*, y de segundo orden, como *causa*).

El subsiguiente proceso de encaje de atributos se encuentra gobernado por dos *restricciones globales*, que garantizan que la colección de hipótesis de encaje resultante es, además de maximal, *estructuralmente consistente*:

(1) *Correspondencias uno-a-uno*: las correspondencias identificadas por el proceso de encaje entre los distintos elementos (a saber, entidades o predicados) deben ser *uno-a-uno*, esto es, cada elemento en uno de los grupos puede ponerse en correspondencia –como máximo– con uno de los elementos del otro grupo.

(2) *Conectividad paralela*[54]: esta condición se cumple cuando los argumentos de los predicados puestos en correspondencia pueden, a su vez, ponerse en correspondencia entre sí. Esta asunción es clave, pues garantiza que el mapeo obtenido no está solamente basado en la similaridad de esos rasgos, sino en el papel que cumplen en la estructura relacional de sus grupos.

Finalmente, el proceso de encaje consistiría en un algoritmo que avanza en la dirección *de-local-a-global*, pasando por cada una de las cuatro etapas siguientes: (i) identificación de encajes locales entre todos los predicados que sean similares (por ejemplo, *Führer-de* y *presidente-de*; u *ocupar* e *invadir*); (ii) integración de los encajes locales identificados en núcleos –o clusters– estructuralmente consistentes; (iii) unión de esos núcleos en un reducido número de conjuntos que mantengan las correspondencias uno-a-uno y sean estructuralmente consistentes, al tiempo que maximales en tamaño; y (iv) selección de uno de dichos conjuntos en función de la métrica escogida (por ejemplo, la métrica podría favorecer mapeos profundos que pusieran en correspondencia relaciones de orden superior).

3.5.3 Modelo SIAM

El *modelo SIAM* (Goldstone 1994a), por su parte, constituye una aproximación de tipo conexionista a los modelos basados en alineamientos. En este caso el modelo consiste en una red neuronal cuyos nodos representan hipótesis con respecto a las correspondencias

[54] Aunque Falkenhainer y colaboradores (1989, §3.2.2) utilizan el término *soporte*, considero preferible la expresión *conectividad paralela* empleada por Hodgetts para referir a la segunda condición (Hodgetts *et al.* 2009, p. 64).

entre los rasgos u objetos constituyentes de dos escenas. La asunción básica del modelo es que la similaridad se puede determinar mediante un proceso dinámico de activación interactiva de correspondencias entre rasgos, objetos y roles. Sobre esta base, el modelo SIAM opera siguiendo el proceso de búsqueda siguiente[55]:

(1) *Correspondencias entre rasgos* (CeR): identificación de correspondencias entre los rasgos de las dos escenas consideradas.
(2) *Correspondencias entre objetos* (CeO): búsqueda de correspondencias entre objetos que sean consistentes con las CeR previamente obtenidas.
(3) *Realimentación y ajuste*: después, un proceso de activación realimenta al sistema, favoreciendo el encaje –o no-encaje– de rasgos en función de su consistencia con los alineamientos de objetos ya encontrados.

Así es como tendría lugar la influencia mutua entre las CeO y las CeR, a saber: las CeO promueven la activación de CeR y, al mismo tiempo, las CeR influyen en la activación de las CeO. O, en otras palabras, el proceso de búsqueda que subyace al modelo SIAM opera de modo tal que, cuando un cierto rasgo –de un objeto *c*– es puesto en correspondencia con otro rasgo –de otro objeto *d*–, a partir de entonces el modelo favorece la identificación de otras CeR que sean consistentes con el alineamiento entre los objetos *c* y *d*[56]. En este modelo, la activación de los nodos de la red neuronal estaría guiada por las dos directrices siguientes: (i) los nodos consistentes se excitan entre sí; y (ii) los nodos inconsistentes se inhiben los unos a los otros.

En último lugar, los alineamientos obtenidos aumentan la contribución a la similaridad –entre dos escenas– de los rasgos puestos en correspondencia por cada nodo. Por ello, la similaridad entre dos escenas *a* y *b* podría determinarse del modo siguiente:

$$s(a,b)=\frac{\sum_{i=1}^{n}\text{encaje}_i A_i w_i}{\sum_{i=1}^{n} A_i w_i}$$

[55] Obsérvese que el modelo SIAM tiende tácitamente hacia las mismas restricciones globales que asumía el modelo SME –a saber, *correspondencias uno-a-uno* y *conectividad paralela*–. Por un lado, promueve la búsqueda de correspondencias consistentes, y considera que dos nodos son inconsistentes si crean alineamientos de *uno-a-varios* (esto es, si varios elementos de una escena pueden ponerse en correspondencia con un solo elemento de la otra). Por otro lado, el proceso de búsqueda del modelo SIAM favorece tanto la identificación de CeR pertenecientes a objetos ya alineados, como el alineamiento de objetos (u obtención de CeO) con CeR consistentes.

No obstante, el modelo SIAM se diferencia de modelo SME en que no se compromete estrictamente con ninguna de estas condiciones. Así, por ejemplo, el hecho de que no se adhiera de modo estricto a la restricción de *correspondencias uno-a-uno* permite que existan –en el modelo SIAM– grados de correspondencia (lo cual no era posible en el enfoque de "todo o nada" propio del modelo SME). Por esta razón, en el modelo SIAM los encajes desalineados también tienen influencia sobre la similaridad (aunque menor que la que tienen los encajes alineados).

[56] Un rasgo de una CeR asociado a un objeto alineado sería un encaje en su lugar –en inglés, *match in place*, o MIP–; mientras que un rasgo de una CeR asociado a un objeto desalineado sería un encaje fuera de lugar –en inglés, *match out of place*, o MOP–.

En donde n es el número de nodos del sistema; A_i es el nivel de activación del nodo i; w_i es la importancia dada a la dimensión i –pudiendo definirse como $w_i = w_{ia} + w_{ib}$, siendo w_{ix} la importancia de la dimensión i en la escena x–; y los valores encaje$_i$ variarán entre 0 y 1 en función de la similaridad existente entre los rasgos puestos en correspondencia por el nodo i.

3.5.4 Puntos fuertes y débiles

Como ya se ha indicado, la principal ventaja de los modelos de alineamiento es que dan cuenta del carácter estructurado de muchas de nuestras representaciones. Con ello, los modelos basados en alineamientos proporcionan una explicación a los estudios empíricos que muestran que los rasgos en correspondencia para objetos correctamente alineados (de dos escenas) contribuyen más a la similaridad entre esas escenas que los rasgos puestos en correspondencia para objetos con una alineación más pobre (Goldstone 1994a; Love 2000). Además, el modo en que operan los procesos de búsqueda de alineamientos explicaría que la contribución a la similaridad de los rasgos alineados y no alineados –esto es, MIPs y MOPs (véase nota al pie 56 en este capítulo)– aumente con el tiempo de procesamiento (Goldstone y Medin 1994); que un mismo rasgo u objeto contribuya más a la similaridad cuando aumenta la claridad de los alineamientos entre las escenas (Goldstone 1994a); y que en ocasiones la introducción de nuevos rasgos pobremente alineados no contribuya a aumentar la similaridad, sino que la reduzca –en la medida en que dicho nuevo rasgo puede interferir en el proceso de búsqueda de alineamientos adecuados (Goldstone 1996)–.

Y, en lo que respecta a las principales dificultades y problemas de este enfoque, Holyoak (2005) considera que uno de los retos más importantes a los que se enfrentan los modelos basados en la identificación de alineamientos estructurales[57] es que en ellos el conocimiento sobre las representaciones –y, en particular, sobre su estructura– debe ser introducido a mano por el creador del modelo, mientras que en el caso de los seres humanos la formación de dicho conocimiento tiene lugar de modo autónomo.

3.6. Modelos transformacionales

Una cuarta aproximación a la noción de similaridad son los *modelos transformacionales* (también llamados *modelos de distorsión representacional*), en los que la similaridad entre dos objetos se encuentra determinada por la distancia transformacional entre sus representaciones. La noción de *distancia transformacional* difiere de la de distancia geométrica en que en ella la medida de distancia refiere a (o es proporcional a) la complejidad de las transformaciones necesarias (o, alternativamente, a la cantidad de distorsión requerida) para pasar de una representación a otra (Chater y Hahn 1997; Hahn y Chater 1997; Chater y Vitányi 2003; Hahn, Chater y Richardson 2003; Hahn, Close y Graf 2009)[58].

[57] Aunque la discusión de Holyoak se centra en el caso de los modelos de analogía, sus conclusiones son de perfecta aplicación al caso de los modelos basados en alineamientos.

[58] Para una generalización de la noción de distancia transformacional más allá de la comparación entre dos objetos individuales véase Bennett *et al.* (1998).

Así, cuanto más simple sea la transformación requerida para convertir la representación de un objeto *a* en la representación de un objeto *b*, mayor será la similaridad entre los objetos *a* y *b*.

3.6.1 Mapeos y transformaciones

Los modelos transformacionales abordan la cuestión de cómo se puede transformar –o mapear– una representación sobre otra. Obviamente, en caso de que exista una relación estructural entre ambas representaciones el mapeo será más simple (y, por ello, su similaridad mayor), puesto que la transformación conjunta de elementos por bloques será más eficiente que su transformación paso a paso.

El hecho de que estos modelos busquen correspondencias –o mapeos– entre los objetos evaluados (para luego, a partir de ellas, transformar un objeto en otro) es el motivo por el que los modelos transformacionales pueden ser vistos como un cierto tipo de enfoque basado en alineamientos. En un caso (*modelos de alineamiento*) las correspondencias estarían explícitamente dadas en los alineamientos estructurales, mientras que en el otro (*modelos transformacionales*) los mapeos serían implícitos a las transformaciones utilizadas[59]. En consecuencia, no resulta extraño que algunas de las primeras propuestas de tipo transformacional surgieran en paralelo a enfoques basados en analogías y/o alineamientos. Así, por ejemplo, Alvin Goodman (1986, p. 242), en su estudio de las analogías entre particiones y jerarquías, sostiene que el contenido de dos representaciones es similar en la medida en que el contenido de una se pueda derivar del contenido de la otra por medio de substituciones, inversiones u otras operaciones similares[60].

Esta misma línea había sido anticipada en los estudios de Imai (1977) con respecto a cómo los sujetos evalúan la similaridad entre objetos que se diferencian en una o varias transformaciones. En particular, Imai presentaba a los sujetos secuencias de elipses blancas y negras, y les pedía que evaluaran su similaridad. A continuación procedía a caracterizarlas en términos de su distancia transformacional en base a las cuatro operaciones siguientes:

— *Imagen especular* (E): ○●○●○○○●●○○● $\rightleftharpoons$ ●○○●●○○○●○●○
— *Desplazamiento de fase* (F): ○●○●○○○●●○○● $\rightleftharpoons$ ●○●○○○●●○○●○
— *Inversión* (I): ○●○●○○○●●○○● $\rightleftharpoons$ ●○●○●●●○○●●○
— *Cambio de longitud de onda* (O): ○○●●○○●●○○●● $\rightleftharpoons$ ○●○●○●○●○●○●

Sobre esta base, la distancia transformacional entre las secuencias ●●●○●●●○●●●○ y ○○●○○○●○○○●○ sería igual a dos, dado que estas dos secuencias se pueden igualar me-

[59] No obstante, el foco de atención en ambas aproximaciones es distinto pues, mientras que en los modelos de alineamiento el objetivo era establecer *correspondencias consistentes* entre las escenas consideradas, en el caso de los modelos transformacionales el interés principal es la *codificación óptima* de la transformación en un lenguaje –en base a la cual la complejidad de la misma pueda determinarse– (Hodgetts et al. 2009, p. 65).

[60] Más adelante Ullman (1996, pp. 97-104) sostendrá que el alineamiento entre objetos tridimensionales se puede realizar –sin necesidad de descomponer los objetos en partes– mediante transformaciones tales como traslaciones, rotaciones, cambios de escala, y torcimientos topográficos.

diante una inversión y un desplazamiento de fase. La investigación de Imai probaba que la similaridad entre dos secuencias era tanto mayor cuanto menor fuera su distancia transformacional en términos de las operaciones anteriores[61,62]. Además, su estudio también probaba que la similaridad entre secuencias transformables en un solo paso era tanto mayor cuanto mayor fuese el número de operaciones que permitiesen esa transformación. Por ejemplo, la similaridad entre las secuencias ○○●●○○●●○○●● y ●●○○●●○○●●○○ –las cuales son mutualmente transformables en un paso mediante tres operaciones distintas (E, F o I)– era mayor que la similaridad entre las secuencias ○○○○○○○○○○○○ y ●●●●●●●●●●●● –también mutuamente transformables en un solo paso, pero ahora solamente con una operación posible (a saber, I)–.

Por último, otro posible enfoque transformacional sería el de Palmer (1983), quien caracterizaba la similaridad –de formas, movimientos, agrupamientos, etc.– en torno a la noción de invariancia bajo el grupo de transformaciones de similitud (a saber, traslaciones, rotaciones, reflexiones y cambios de escala), esto es, para transformaciones en las que se mantiene invariante la forma, pero no el tamaño ni la posición de los elementos considerados.

3.6.2 Complejidad de Kolmogorov

La complejidad de una transformación –o cantidad de distorsión necesaria – puede ser calculada por medio de la teoría de la complejidad algorítmica (Li y Vitányi 2008), también llamada *teoría de la complejidad* de Kolmogorov (1963, 1965)[63]. Conforme a esta teoría, la complejidad K de un objeto, cadena o representación x es igual a la longitud del menor programa –en un cierto lenguaje de computación– capaz de producir dicho objeto. O, en otras palabras, es una medida de la cantidad de recursos informacionales necesarios para generar una determinada representación. Por ello, las representaciones producidas por programas cortos son consideradas más simples que aquellas generadas por programas más largos. Así, por ejemplo, una serie de un billón de 1s, a pesar de su gran longitud, puede ser producida por un programa muy corto, por lo que su complejidad es muy baja. Solo un poco mayor sería la complejidad de una serie de un billón de 0s y un billón de 1s alternos, y algo mayor aún la de los términos de la serie de Fibonacci menores o iguales que un billón. La siguiente sería la codificación en lenguaje C de los menores pro-

[61] De modo similar, los experimentos de Wiener-Ehrlich *et al.* (1980) también confirmaban la existencia de correlaciones entre los juicios de similaridad de los sujetos y el número de transformaciones necesarias para convertir un estímulo en otro.

[62] Años después Hahn, Chater y Richardson (2003) confirmaron –con un diseño experimental basado en el trabajo de Imai– la significancia estadística de las correlaciones existentes entre similaridad y distancia transformacional para este tipo de secuencias.

[63] Para profundizar en la importancia y justificación de la simplicidad –y, por consiguiente, de la medición de la complejidad– en computación y ciencia cognitiva, así como de sus aplicaciones, véanse Wallace y Boulton (1968), Rissanen (1978, 1989), Quinlan y Rivest (1989), Chater (1996, 1999), así como Wallace y Dowe (1999).

gramas capaces de producir las cadenas anteriores, junto con sus complejidades de Kolmogorov asociadas en ese lenguaje[64, 65]:

— *Serie de un billón de 1s* (*S1*):

1 1 1 1 1 1 1 1 1 1 ...

```
main(){for(long long i=1;i<=pow(10,12);i++){printf("1 ");}}
```

$K_{\mathrm{C}}(\mathit{S1})=58$

— *Serie de un billón de 0s y 1s alternos* (*S01*):

0 1 0 1 0 1 0 1 0 1 0 1 ...

```
main(){for(long long i=1;i<=pow(10,12);i++){printf("01 ");}}
```

$K_{\mathrm{C}}(\mathit{S01})=59$

— *Serie de Fibonacci con términos menores o iguales que 1 billón* (*SF*):

0 1 1 2 3 5 8 13 21 34 55 89 ...

```
main(){long long i=0,j=1,a;while(i<=pow(10,12)){printf("%d ",i);
a=j;j=j+i;i=a;}}
```

$K_{\mathrm{C}}(\mathit{SF})=80$

En cambio, el texto de *El Quijote*, aunque mucho más corto que un billón de caracteres[66], no puede ser generado por un programa sencillo como los anteriores, razón por la cual su complejidad es mucho mayor. Finalmente, la complejidad de una cadena de números aleatorios es la mayor posible –e igual a la longitud de la propia cadena–, pues su menor descripción en este caso es una mera copia literal de la cadena aleatoria[67].

Ya de manera general, la complejidad de Kolmogorov puede definirse del modo siguiente. Sea A^* el conjunto de palabras que pueden formarse con un cierto alfabeto A, en cuyo caso $\{0,1\}^*$ representará el conjunto de toda posible cadena binaria[68]; y sea una función recursiva parcial $\varphi : \{0,1\}^* \rightarrow \mathcal{O}$, donde $\mathcal{O} \subseteq \{0,1\}^*$. La cadena y es una descripción de la cadena x si se cumple que $\varphi\,(y) = x$, esto es, si la función φ produce como resultado x cuando se le introduce como entrada la cadena y. Sobre esta base, puede definirse la complejidad de Kolmogorov ($K_\varphi : \mathcal{O} \rightarrow \mathbb{N}$) de una cierta cadena x con respecto de φ, de la manera siguiente:

$$K_\varphi(x) = \min\{|\,p\,| : \varphi(p) = x\}\,, \quad (\text{o } \infty \text{ si } p \text{ no existe})$$

[64] Los programas considerados separan cada uno de los elementos de estas series por medio de espacios.

[65] En todos estos programas las variables empleadas necesitan ser definidas como tipo `long long` –y no meramente como tipo `int`– con objeto de que puedan soportar números enteros hasta el billón.

[66] La longitud aproximada de *El Quijote* son dos millones de caracteres.

[67] Obsérvese que ése no era el caso para el texto de *El Quijote*, cuya longitud (debido a la redundancia y diferente frecuencia de las letras y palabras del castellano) puede reducirse –mediante algoritmos de compresión de texto– a menos de setecientos mil caracteres.

[68] El conjunto $\{0,1\}^*$ es, por tanto, susceptible de poder ser identificado con el conjunto de los números naturales $\mathbb{N}$, conforme a la correspondencia siguiente: (0,ϵ), (1,0), (2,1), (3,00), (4,01), (5,10), etc., en donde ϵ representa la cadena vacía.

En donde $|p|$ representa la longitud de la cadena p.

La expresión anterior puede interpretarse del modo siguiente. Si p es un programa, φ es una estructura capaz de ejecutar programas (por ejemplo, φ podría ser la combinación de un lenguaje de programación, más su compilador, más el hardware preciso para ejecutar sus programas compilados[69,70]), y $\varphi(p)$ es la salida generada por la ejecución del programa p –siendo $\mathcal{O}$ el conjunto de todas las salidas posibles–; entonces, para una cierta cadena $x \in \mathcal{O}$, la complejidad de x con respecto a φ (esto es, $K_\varphi(x)$) es igual a la longitud del programa más corto p que permite computar x en φ (esto es, $\varphi(p)=x$).

La complejidad de Kolmogorov se puede extender al ámbito del cálculo de similaridades entre dos objetos a y b mediante lo que se conoce como *complejidad condicional de Kolmogorov*, la cual se define como la longitud del menor programa p capaz de transformar la representación de a en la representación de b (o, dicho de otro modo, capaz de generar la cadena a utilizando la cadena b como información auxiliar de entrada[71]):

$$K_\varphi(a\,|\,b) = \min\{|p|: \varphi(p,b)=a\}, \quad (\text{o } \infty \text{ si } p \text{ no existe})$$

Por consiguiente, la similaridad entre dos cadenas a y b estaría determinada por el número y longitud de las instrucciones necesarias para transformar la una en la otra, siendo tanto mayor –la similaridad– cuanto menos compleja resulte la transformación. Así, por ejemplo, las secuencias 1 2 3 4 5 6 7 8 y 2 3 4 5 6 7 8 9 serían muy similares, en la medida en que se requiere de una sola instrucción simple (a saber, *añadir/substraer 1 a cada dígito*) para transformar la una en la otra. Por la misma razón, las secuencias 1 2 3 4 5 6 7 8 y 2 4 6 8 10 12 14 16 serían igualmente similares (en este caso la instrucción sería *multiplicar/dividir por 2 cada dígito*). En cambio, la similaridad entre las secuencias 1 2 3 4 5 6 7 8 y 3 7 9 11 13 15 17 sería ya menor, pues la transformación de la una en la otra precisa de dos operaciones (a saber, *multiplicar por 2 y sumar 1*, y *restar 1 y dividir por 2*, respectivamente). A continuación se presenta la codificación en Bash Shell de los tres pares de transformaciones anteriores, junto con sus correspondientes complejidades (condicionales) de Kolmogorov en ese lenguaje:

— Transformación 1 2 3 4 5 6 7 8 (a) $\rightleftharpoons$ 2 3 4 5 6 7 8 9 (b):

```
for i in $1;do echo $((i+1));done
for i in $1;do echo $((i-1));done
```

$K_{\mathrm{BS}}(a\,|\,b)=33$

$K_{\mathrm{BS}}(b\,|\,a)=33$

— Transformación 1 2 3 4 5 6 7 8 (a) $\rightleftharpoons$ 2 4 6 8 10 12 14 16 (b):

```
for i in $1;do echo $((i*2));done
for i in $1;do echo $((i/2));done
```

$K_{\mathrm{BS}}(a\,|\,b)=33$

$K_{\mathrm{BS}}(b\,|\,a)=33$

— Transformación 1 2 3 4 5 6 7 8 (a) $\rightleftharpoons$ 3 7 9 11 13 15 17 (b):

```
for i in $1;do echo $((i*2+1));done
```

$K_{\mathrm{BS}}(a\,|\,b)=35$

[69] En mis anteriores ejemplos he referido a todo esto abreviadamente como "lenguaje C".

[70] De manera general, φ podría ser cualquier máquina de Turing.

[71] Obviamente, $K_\varphi(x)$ es un caso particular de complejidad condicionada, en donde la información auxiliar de entrada es la cadena vacía, esto es, $K_\varphi(x\,|\,\epsilon)$.

```
for i in $1;do echo $(((i-1)/2));done
```

$K_{\mathrm{BS}}(b|a)=37$

3.6.3 Ventajas del enfoque transformacional

La principal ventaja de los modelos basados en transformaciones es que, como ya se ha visto, permiten dar cuenta de fenómenos no explicados por los otros modelos de similaridad. Por un lado, los modelos transformacionales explican por qué las personas consideran que dos objetos son tanto más similares cuanto menor es el número de transformaciones necesarias para convertir el uno en el otro (Imai 1977; Wiener-Ehrlich *et al.* 1980; Hahn, Chater y Richardson 2003; Hodgetts 2011). O, dicho de otro modo, la distancia transformacional explica la puntuación de similaridad dada por los sujetos.

Además, este tipo de enfoques también permite explicar el hecho de que para pares de series transformables en un mismo número de pasos *n*, la similaridad entre ellas varíe y lo haga inversamente al número de operaciones que permiten dicha transformación (Imai 1977). En este caso la idea es que, cuanto mayor sea la variedad de transformaciones que conducen de un objeto a otro, más fácil –y rápido– será que el sujeto encuentre una de ellas y, por ello, mayor será la similaridad entre ese par de secuencias.

Por último, el hecho de que las propuestas transformacionales modelen la similaridad entre dos objetos como una medida inversa de la complejidad de la transformación precisa para convertir el uno en el otro[72], explica la variación de la similaridad en función de lo complejas que sean las operaciones intervinientes en la transformación, así como su posible carácter asimétrico –en caso de que la complejidad de las operaciones precisas para las transformaciones en uno y otro sentido sea distinta– (Hahn, Close y Graf 2009). En este último caso, las asimetrías en la complejidad de las transformaciones explicarían la existencia de asimetría en los juicios de similaridad. Asimismo, esta relación entre complejidad y similaridad estaría respaldada por estudios que muestran la existencia de una correlación positiva entre la complejidad de una transformación y el tiempo necesario para su reconocimiento[73] (Tarr y Pinker 1989; Cave *et al.* 1994; Tarr 1995).

3.6.4 Problemas y limitaciones

En cuanto a las limitaciones de los modelos transformacionales, uno de sus puntos débiles es que se trata de propuestas que han sido casi exclusivamente aplicadas al caso de estímulos perceptuales, estando en el aire su capacidad para emplearse también para estímulos conceptuales –o, cuando menos, su capacidad para hacerlo sin convertir en el proceso al modelo transformacional en un modelo de alineamiento– (Goldstone y Son 2005, p. 27). Esto, unido al hecho de que los modelos de alineamiento hayan sido aplicados –ante

[72] Esto era una extensión natural de la asunción –común en los modelos transformacionales– de que el número de operaciones necesarias para convertir un objeto en otro resulta indicativo de la complejidad de la transformación. Obsérvese que, bajo dicha asunción, una vez postulado un cierto repertorio operacional –esto es, el conjunto de transformaciones básicas–, la complejidad transformacional permite predecir las similaridades evaluadas (Hahn, Chater y Richardson 2003).

[73] A lo que acompaña una relación inversa entre el tiempo de respuesta y la práctica, lo que sugiere que la experiencia contribuye al desarrollo de mecanismos que simplifican –o agilizan– el procesamiento de estímulos complejos.

todo– a inputs conceptuales (y que no sean capaces de explicar algunos fenómenos sí explicados transformacionalmente) apunta a que una prometedora línea de investigación sería la combinación de ambos enfoques, tal y como hace la arquitectura Copycat (Mitchell, 1993; Hofstadter y Mitchell 1994; Hofstadter 1995).

Otra dificultad de las propuestas basadas en transformaciones es que la articulación psicológica de estos enfoques requiere que se especifique un cierto lenguaje que codifique las transformaciones, y en el cual la longitud del código del menor programa capaz de realizar una cierta transformación determine la complejidad de la misma. El problema es que la opción por un lenguaje concreto para las transformaciones es una decisión que nunca estará libre de controversia.

Finalmente, estos modelos asumen una relación entre la similaridad entre dos objetos y la complejidad de la transformación que convierte al uno en el otro que, en algunas ocasiones, presenta un comportamiento indeseado. En este caso el problema es que al aumentar la complejidad de los objetos considerados aumentará la complejidad de la transformación que los relaciona (por el mero hecho de ser objetos más complejos) y, por consiguiente, la distancia transformacional entre ambos. La consecuencia es que, a igualdad del resto de factores, pares de objetos más complejos serán considerados más disimilares entre sí –que otros pares de objetos que fuesen más simples– por el simple hecho de que su complejidad de Kolmogorov sea mayor (Hahn, Chater y Richardson 2003, p. 4). No obstante, a pesar de lo anti-intuitivo que resulta este efecto, su impacto sobre este tipo de modelos es limitado, pues se podría evitar normalizando la complejidad de las transformaciones por la complejidad de los objetos considerados (Bennett *et al.* 1998; Vitányi *et al.* 2009).

3.7. Recapitulación

En el capítulo anterior ya comenté las principales objeciones planteadas a los enfoques basados en similaridades, y entre las que destacaban los problemas de composicionalidad[74] y selección. Allí también se apuntaba a que ciertos investigadores sugerían que la noción de similaridad no era lo suficientemente flexible como para explicar la cognición, razón por la cual se proponían modelos alternativos, basados en reglas (Smith y Sloman 1994) y en teorías (Murphy y Medin 1985). No obstante, y aún cuando –conforme indicaba al final del capítulo 2– en mi opinión es necesario considerar otros elementos (tales como teorías y definiciones) a la hora de explicar muchos fenómenos cognitivos, eso no significa que la similaridad no tenga un papel siempre presente, tal y como muestra el hecho de que los sujetos tengan problemas para ignorar patrones similares aún cuando disponen de reglas de categorización absolutamente precisas (Palmeri 1997).

Por otro lado, una crítica indirecta a los modelos de similaridades sería la de Fodor (2000) cuando, sobre la base de que la cognición humana se caracteriza por su sensibilidad al contexto –siendo ésta una tesis comúnmente aceptada–, se muestra escéptico en cuanto a que dicha sensibilidad pueda explicarse mediante las actuales teorías com-

[74] Para una discusión de cómo las similaridades podrían componer para un sistema de "conceptos naturales" –en el sentido dado por Gärdenfors (2000) a la noción "concepto natural"– véase Leitgeb (2005).

putacionales de la mente. En este caso Fodor entiende que un sistema representacional con estructuras causales fijas no puede operar de modo sensible al contexto, lo cual pone en cuestión la posibilidad de que se pueda explicar computacionalmente la sensibilidad al contexto de la cognición (y, por consiguiente, de la similaridad)[75]. En respuesta a esta objeción se ha argumentado que es posible implementar sistemas computacionales simples que muestran que la estructura de similaridad de una representación puede cambiar con el contexto, para redes neuronales tanto simples (Thomas *et al.* 2012) como complejas (Thomas y Mareschal 2001; Rogers y McClelland 2004). La idea en todos estos casos es que la cognición puede operar sobre un mecanismo de modulación conceptual que permitiría que representaciones invariantes al contexto generen representaciones sensibles a ese mismo contexto.

No obstante, y a pesar de todas esas objeciones, los modelos de similaridades tienen como principal ventaja el que gozan de una gran capacidad explicativa sobre la base de una asunción naturalista básica. O, en otras palabras, las teorías de similaridad constituyen un medio para explicar fenómenos tales como la adquisición de conceptos, producción de generalizaciones, categorización de objetos, realización de inferencias, etc., y lo hacen en base a la asunción de que los objetos y eventos similares se comportan de modo similar en circunstancias semejantes. El principal punto fuerte de esta asunción es que resulta poco controvertida, en la medida en que hay buenos motivos evolutivos para su aceptación pues, si entidades similares no se comportasen de modo semejante –en contextos parecidos– sería difícil concebir, no solo cómo ninguna especie podría haber llegado a adquirir concepto alguno, sino incluso sobre qué base podrían operar los mismos procesos de selección natural.

En este capítulo he presentado cuáles son los principales modelos contemporáneos de similaridad, junto con sus principales puntos fuertes y débiles. También he puesto de manifiesto que ninguno de tales modelos es individualmente capaz de explicar todos los distintos fenómenos empíricos identificados, lo que apunta a la posible conveniencia de una postura de tipo pluralista, abierta a la inclusión de elementos de los diferentes modelos (tal y como ocurría para el problema de la estructura de los conceptos). En todo caso, de entre las distintas teorías de similaridad presentadas me inclino porque sea un enfoque de tipo geométrico el que actúe como núcleo en torno al que luego desarrollar una propuesta pluralista más amplia. Por ello, el resto del presente trabajo asumirá una aproximación geométrica a la idea de similaridad.

Conforme se ha visto en este capítulo, los modelos geométricos conciben la similaridad entre objetos como proximidades espaciales, esto es, como una función inversamente proporcional a las distancias en un espacio métrico. Sobre esa base, los modelos de tipo geométrico representan los objetos –o entidades– como puntos localizados en un espacio organizado en dimensiones y construido, por ejemplo, a partir de juicios de (di)similaridad. En cuanto a esto, el escalamiento multidimensional ha sido histórica-

[75] Aunque el problema de la sensibilidad al contexto ha sido presentado en este capítulo para algunos modelos concretos de similaridad, esas críticas específicas ya vistas se diferencian de la objeción de Fodor en que ésta resulta mucho más general, en la medida en que no cuestiona un modelo de similaridad concreto, sino la posibilidad de caracterizar la similaridad mediante un sistema computacional con estructuras fijas.

mente la aproximación preferida para la construcción de modelos geométricos a partir de las (di)similaridades observadas entre cada par de elementos de un cierto conjunto de objetos, dada su capacidad para encontrar representaciones del conjunto de objetos inicial en espacios con dimensionalidad reducida (esto es, con un menor número de dimensiones que el espacio original), y en los que las distancias entre cada par de objetos encajan lo más posible con las distancias originalmente existentes entre ellos. Finalmente, aún y a pesar de las críticas recibidas por la violación de sus axiomas, este tipo de modelos presentaba notables ventajas, entre las que destaca (a) su capacidad para identificar las dimensiones subyacentes a un conjunto inicial de datos, (b) la producción y manejo de espacios con dimensionalidad reducida –lo cual era muy conveniente en términos de codificación, memoria y procesamiento–, y (c) su posible uso para caracterizar funciones cognitivas tales como categorizaciones, inferencias y memoria, etc.

En mi caso, la razón que justifica mi preferencia por los modelos geométricos es su carácter concreto e integrador, pues constituyen un planteamiento que integra de modo coherente en una única teoría los principales elementos y problemas presentes en las discusiones relativas a la naturaleza y formación de conceptos. En primer lugar, el modelo geométrico proporciona el mismo marco para la representación de objetos y conceptos en el que, además, los atributos –o dimensiones– se integran de modo natural (esto es, como un elemento propio del modelo). Por otro lado, explica cómo los conceptos pueden formarse a partir de objetos –o eventos– particulares, pudiendo dar cuenta también de cómo puede tener lugar la adquisición de conceptos a partir de información perceptual, aún cuando la cuestión de cuál es la naturaleza y origen de esas primeras dimensiones siga siendo problemática.

Además, como veremos en los capítulos siguientes, un modelo geométrico articulado mediante espacios de similaridad conceptual facilita la explicación de importantes procesos cognitivos, tales como: (a) *memoria* –en el sentido de almacenaje– la cual es más factible cuando se la concibe operando sobre espacios con dimensionalidad reducida como los producidos por los modelos geométricos de similaridad; (b) *aprendizaje*, entendido como subdivisión –o composición– recurrente de regiones del espacio de similaridad, lo que da lugar a la producción de conceptos más –o menos– específicos que los existentes; y (c) *inferencias*, tanto *deductivas*, cuando la representación de un cierto objeto/concepto está incluida en la región asociada a otro concepto, como *inductivas*, cuando la formación de un nuevo concepto tiene lugar a partir de la información de objetos particulares.

Chapter 4: Conceptual space theories

[L'universo] è scritto in lingua matematica, e i caratteri son triangoli, cerchi, & altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola. –
Galileo Galilei (1623, VI, p. 25)

The notion of *conceptual space* –as a framework for the representation of concepts and knowledge– has been highly influential over the last decade or more, as a way of articulating the geometric model of similarity that avoided the difficulties besetting the geometrical perspective. The first half of this chapter is devoted to showing what a similarity space theory of concepts is and, in particular, to describing Gärdenfors' conceptual spaces and the role played by convexity in this latter approach. In the second part I will argue that Gärdenfors' convexity constraint on the shape of regions is both unnecessary –from a theoretical perspective– and problematic –with regard to some particular aspects of how the conceptual space theory works–. Lastly, I conclude that if the convexity condition is abandoned, then Gärdenfors' theory of conceptual spaces is equivalent to a contextualist geometric characterization of the prototype theory of concepts.

As said in the previous chapter, the *standard* geometric model of similarity is subject of criticism for its inability to explain the context-sensitivity of similarity judgments, and in particular, for the violations of the metric axioms (i.e., minimality, symmetry, and triangle inequality) assumed by this sort of model. One possible way of responding to this set of objections was to adopt a *contextualized* geometrical notion of similarity, and that is precisely the perspective embraced by the conceptual space theories here discussed.

With this goal in mind, the first section of this chapter aims to describe the (similarity-based) space theories of concepts. There I introduce the main notions and principal theses of the theory, together with the major approaches to the theories of conceptual spaces (i.e., connectionist versus geometrical). I also recap how the notion of similarity is characterized within the geometrical approach, and outline the distinction between standard and non-standard distances. Then, after introducing the notion of Voronoi partition, I sketch out the most significant strengths of the geometric view on conceptual spaces.

In the second section I expound the main features of Gärdenfors' theory of conceptual spaces, focusing on his definition and characterization of properties and concepts

(Gärdenfors 2000, 2014). One of the basic theses of Gärdenfors' approach is that the conceptual regions associated with properties, concepts –or object categories–, verb meanings[1], etc. are convex. At that point I will explain the role played by the convexity requirement in the theory, as opposed to other possible criteria that could be imposed on the geometry of regions. Lastly, I also summarize some of the most remarkable applications and extensions of Gärdenfors' conceptual spaces.

My aim in the third section is to show that the convexity constraint –according to which the geometry of conceptual regions should be convex– is questionable. On the one hand, I will show that all the *direct* arguments given by Gärdenfors in favor of the convexity of conceptual regions rest on controversial assumptions. Additionally, I will hold that his argument in support of a Euclidean metric –based on the integral character of conceptual dimensions– is weak, and under non-Euclidean metrics the structure of regions may be non-convex. Furthermore, I will prove that, even if the metric were Euclidean, the convexity constraint could be not satisfied if concepts were differently weighted. On the other hand, I will claim that Gärdenfors' convexity requirement is brought into question by some applications of the conceptual spaces theory: (i) several of the allegedly convex properties of concepts are not convex; (ii) the conceptual regions resulting from the combination of convex properties can be non-convex; (iii) convex regions could covary in a non-convex way; and (iv) his definition of changes –linguistically expressed by means of verbs– is incompatible with a definition of properties in terms of convex regions. Then I claim that once the convexity constraint is given up to, the conceptual space theory may be viewed as a contextualist geometric articulation of the prototype theory of concepts.

Consequently, even though the rest of my thesis will be based, to a greater or lesser extent, on the conceptual space theory, I will diverge in significant ways from Gärdenfors' view. First of all, I bring into question the mandatory character of the convexity requirement for the geometry of regions in a similarity space theory of concepts, and I do not adhere to any other compulsory condition on the shape of conceptual regions. Secondly, my starting point for chapters 5 and 6 will be, not Gärdenfors theory of conceptual spaces, but a contextualist geometric view of the prototype theory. On this basis, and in order to respond to some important critiques to the theory, in chapter 5 I will delve into issues rarely addressed by the conceptual spaces literature. There, and after shifting the focus from conceptual regions to prototypes[2], I claim that two different notions of concept should be distinguished (i.e., *stored concept* and *instantiated concept*), which has consequences both for the cognitive/computational architecture presumed by the theory, and for the ontological status attributed to the idea of concept. Those ontological implications are discussed when I argue that, if concepts are assumed to be context-dependent, then concepts lack persistence and do not have representational character –that is, cannot be a representation of their associated categories–. Lastly, in chapter 6, I will deal

[1] Given that since his initial proposal, Gärdenfors (2014) has tried to extend his framework both to the modeling of actions and events, and to the semantics of verbs, prepositions and adverbs.

[2] This is again an explicit movement away from Gärdenfors' theory, since he uses to identify properties and concepts with (convex) *regions*.

with the circularity problem, which threatens, not only the conceptual space theory, but also any empiricist theory of concepts when trying to explain how the most basic elements of concepts can be acquired.

4.1. Similarity space theories of concepts

The prototype theory of concepts –also called *probabilistic view* (Medin 1989) or *family resemblance view* (Komatsu 1992)–, already introduced in chapter 2, maintains that concepts may be organized around sets of correlated attributes that shape an ideal representation –called *prototype*– which sums up the characteristic properties of the considered category. By virtue of this, it is usually said that prototypes are representations –or bodies of knowledge– whose structure encodes information about the properties that the members of a given category tend to have. However, there are distinct ways in which the prototype theory can be articulated (Smith and Medin 1981):

(a) *Featural models*: an object *o* is classified under a concept *C* if it possesses a sufficient number of the properties associated to *C*.

(b) *Dimensional models*: an object *o* is classified under a concept *C* if it *possesses to some degree* a sufficient number of those properties[3].

In both cases an object *o* will be categorized or not under a particular concept *C* in function of the similarity between *o* and the prototype of *C*, which will be determined by virtue of their shared properties. If, for the more general case of dimensional models, the objects and the prototypes of concepts were represented in a geometrical space whose dimensions were the constitutive properties of those concepts –in the relevant context–, then that would be what is commonly called a *similarity space theory of concepts*. In these theories, concepts are located within similarity spaces where distances between concepts and/or objects are inversely proportional to the similarity existing between them (Churchland 1990; Gärdenfors 2000).

4.1.1 Main notions and principal thesis

In general terms, a *similarity space theory of concepts* can be described by one fundamental thesis (Gauker 2007): the mind is a representational hyperspace within which (a) *dimensions* –or *factors*– f_i represent ways in which objects can differ, (b) *points* p_j represent objects, (c) *regions* R_K represent concepts, and (d) *distances* $d_{u,v}$ are inversely proportional to similarities –between objects or concepts–. Consequently, an object *o* will belong to a concept *C* if and only if the values of *o* in every dimension of that similarity space produce an *n*-tuple that lies inside the region associated with the concept *C*.

For instance, Fig. 4.1 shows a conceptual similarity space constituted by *n* dimensions f_i, where the concepts *A* and *B* are represented by the regions R_A and R_B. The points p_j represent distinct objects, three of which (p_1 to p_3) are categorized under the concept *A*, while the other four (p_4 to p_7) are categorized under the concept *B*. The similarity be-

[3] In fact, featural models are nothing more than a particular case of dimensional models.

tween two objects –p_3 and p_7, for example– would be inversely proportional to the distance between them ($d_{3,7}$).

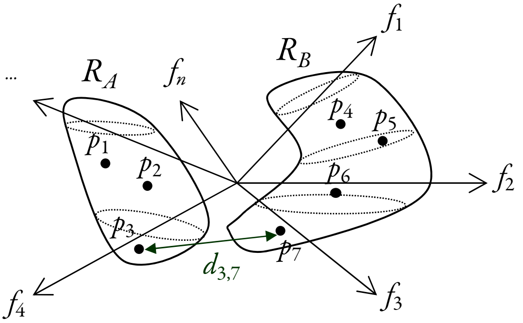

Fig. 4.1. Illustrative example of a conceptual similarity space.

The prototypes of concepts[4] would result from a process of maximization of similarities –or, alternatively, minimization of distances– between the evaluated objects, and the tentative prototypes. The set of final prototypes will be the one which maximizes intragroup similarity and minimizes inter-group similarity. Thence, the prototype of a concept arises as the generalization of the properties of the objects chosen as tentative members of its associated category –for instance, by means of the average of the values in each dimension of the considered objects– (Reed 1972; Hintzman 1986; Nosofsky 1986). Thus, the prototype of a concept would be the most typical member of that category, and would be represented by a point p_p which may correspond or not with a real instance of such category. Lastly, as I will show in section 4.1.3, the shape and boundaries of the conceptual regions may result from a Voronoi tessellation of the conceptual space, whose input are the prototypes of the set of relevant concepts[5].

Inasmuch as distances –or similarities– are a function of variables and parameters which might depend on context (e.g., the relevant concepts, the kind of metric, or the importance of dimensions and concepts), each new instantiation of a concept in a particular context may be different. On this basis, a prototype theory of concepts –con-

[4] In regard to the expression "prototypes of concepts", it might be understood as that "concepts are prototypes" or as that "concepts have prototypes". The first reading is common between those cognitive scientists in favor of the prototype theory of concepts (Rosch and Mervis 1975; Smith *et al.* 1988; Smith and Minda 2002). The second interpretation may be attributed those who think that, although concepts have prototypes, they are other kind of thing –as happens in the case of Gärdenfors, when he identifies concepts with sets of convex regions–. For my part, I postpone further discussion of this issue until section 5.2.2, where this question will be examined in detail.

[5] Nevertheless, prototypes and conceptual regions are two very different things and, as I argue in chapter 5, concepts should be identified with the prototypes –and not with the conceptual regions–, due to two main reasons: (a) what results from the generalization of a set of tentative examples of a given category is a prototype –not a region–; and (ii) in order to categorize an object only the locations of the relevant prototypes are needed (Hernández-Conde 2017a).

ceived in terms of similarity-based spaces– can provide a successful characterization of the contextualist approach to cognition and concepts[6].

4.1.2 Two possible approaches: connectionist versus geometrical

As explained by Gauker (2011), the similarity space theory of concepts emerged from the methods of multidimensional scaling developed by Torgerson (1952) and Shepard (1962a) in order to represent data in terms of their comparative similarity. This kind of techniques were later applied to semantics, as a method used for the identification and modeling of concepts –characterized as regions within a similarity space–, and in categorization research (Nosofsky 1986; Shepard 1987), which finally led to the identification of concepts with regions in a mental space.

However, the first greatly detailed development of a similarity space theory of concepts had to wait until the work of Churchland (1990), which happened with the turn of cognitive science towards connectionism. Ten years later Gärdenfors (2000) proposed a non-connectionist conceptual space theory free from the specific criticisms received by Churchland's proposal due to its connectionist character. Thus, at least two main streams may be distinguished within the similarity-based space theories of concepts:

— *Connectionist approach*: under this view the mind is conceived as a three-layered network, whose associated conceptual spaces are determined by the dimensions which represent the activation levels of the hidden units of the network (Churchland 1990, 1995). Those hidden units constitute the second level of a connectionist network, whose activation levels allow to classify external stimuli –codified by the input units in the first level of the network– into a particular concept or category[7] –determined by the output units in the third level of the connectionist network– (see Fig. 4.2).

  This approach to similarity space theories was severely criticized by Fodor and Lepore (1992), who held that there was no reason in Churchland's proposal for thinking that the hidden unit activation spaces –where Churchland placed concepts– were the same for each particular subject. Or, in other words, Fodor and Lepore's criticism was that the connectionist approach cannot explain the likeness between the *same* concept C for two different subjects S_1 and S_2, since it is not possible to be sure that the activation spaces of S_1 and S_2 are the same, nor consequently that C has the same position –or proximate positions– in them[8]. However, Fodor and Lepore's challenge is not a specific problem of the connectionist view, but a general phenomenon that also applies to other subsequent characterizations of the similarity space theory of concepts, like that of Gärdenfors[9].

[6] All the issues mentioned in this paragraph will be explained in detail in chapter 5.

[7] These concepts or categories could be identified with the regions of the conceptual hyperspace constituted by the activation spaces of the hidden nodes.

[8] In order to know Churchland's response to Fodor and Lepore's critique see Churchland (1998).

[9] Ultimately, since the internal configuration of a neural network is the result of training from a set of external stimuli that may be distinct for each particular subject, the hidden-unit activation spaces –and

Another problem for the connectionist view is that, in a three-layered network, the structure of the hidden-layer can only be determined –or discovered– by an external observer[10], so that representational level is not accessible for other mental processes, which poses significant problems for the characterization of high-level cognition (Halford 2005, p. 536). Finally, another related difficulty of the connectionist approach –associated with how the neural networks codify conceptual information– is its low explanatory capacity regarding the ordinary descriptions of particular concepts (e.g., DOG, CHAIR, etc.).

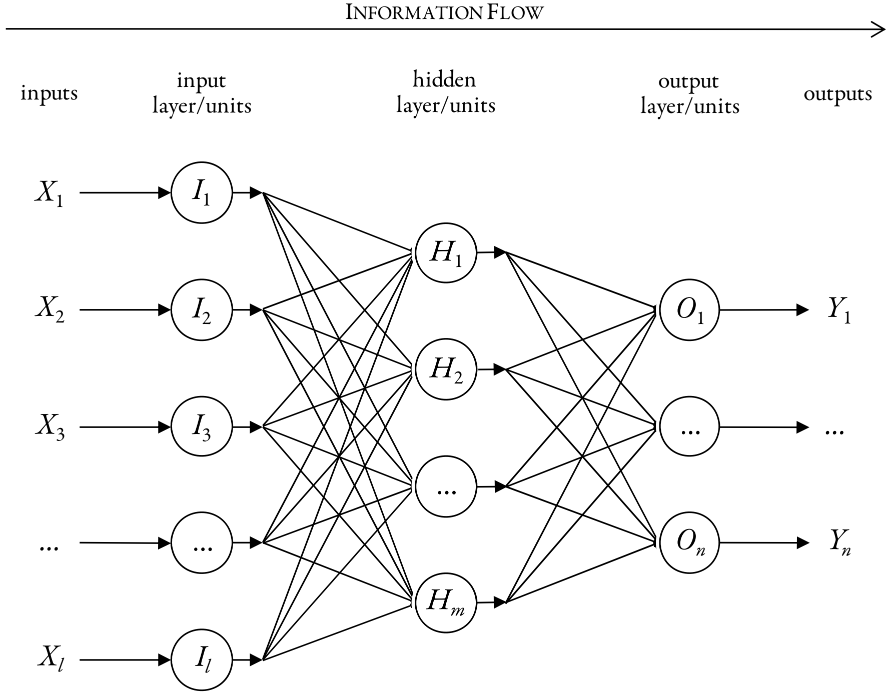

Fig. 4.2. Illustrative example of a simple connectionist network.

— *Geometrical approach*: in contrast, Gärdenfors (2000, 2014) proposes a non-connectionist similarity space theory of concepts, based on the hypothesis that concepts can be located through the combination of quality dimensions (i.e., sensorial properties, which play the role of innate features of our cognitive system) and oth-

therefore their associated concepts– could differ from one individual to another. Anyhow, as said in chapter 1, this kind of difficulty (i.e., that distinct experiential inputs and biographies surely produce different concepts) has to be faced by any empiricist approach to concepts.

[10] By means of, for instance, other analysis techniques such as data clustering (Elman 1990a).

er learned dimensions. According to Gärdenfors' view[11], concepts result from the partition of the similarity space into *convex regions*, which are identified with concepts, constituted by the sets of points representing those objects which exhibit the key sensory properties associated to each considered region (see Fig. 4.1).

One significant advantage of the geometrical perspective over the connectionist view, possible the main one, is that its codification of conceptual information is far more easily interpretable than the results of an approach based on artificial neural networks[12] (as is the case of Churchland's proposal).

Additionally, Gärdenfors also seems to think that his conceptual spaces are free from the threat of Fodor and Lepore's objection regarding the inability of this sort of theories to explain how different subjects may share the same concepts. According to him, the reasons why distinct individuals are able to agree on projectible properties are (i) that those properties are closely tied to the information provided by our senses; and (ii) that many of those (sensorial) quality dimensions are *innate* (Gärdenfors 2000, pp. 26-30). However, such a conclusion is questionable, since nothing guarantees the non-existence of learned factors in the set of dimensions that constitute a given projectible property. And, if other learned factors existed, then those projectible properties might depend on the personal experience and biography of each particular subject, so Fodor and Lepore's challenge would also apply to the geometrical view (see footnote 9 in this chapter).

Once made the observations above, the rest of my thesis will be mainly focused on the geometrical approach and, in the particular case of the present chapter, on Gärdenfors' conceptual spaces.

4.1.3 Similarity measures and Voronoi diagrams

Since in a similarity-based space theory of concepts similarities are computed through distances, and the classification of an object under a particular category is determined by the partition of a conceptual space into Voronoi cells, the aim of this subsection is to briefly introduce the notions of *similarity measure* and *Voronoi diagram*.

[11] Although Gärdenfors' theory of conceptual spaces is the most notable and developed geometrical model within the similarity-based approaches, it is not the only game in town. For instance, other similarity-based geometrical proposals are the *quality spaces* stood by Austen Clark (1993, 2000) –based on the work of Goodman (1951)–, and the *sensory similarity spaces* suggested by Matthen (2005). Anyhow, there are significant reasons why Gärdenfors' view is preferable to Clark's and Matthen's ones. First, Gärdenfors' approach is more explicitly contextual, and –as said in chapter 3– context is critical in order to explain a high number of cognitive phenomena. Second, Gärdenfors' view is centered in the case of concepts, in contrast with those of Clark and Matthen which are focused on the sphere of sensation-perception. (Or, in other words, Gärdenfors' spaces are *conceptual*, while Clark's and Matthen's spaces are *sensorial-perceptual*.)

[12] The difficulty of interpreting (artificial) neural networks is the reason why they are commonly described as black box models.

DISTANCES AS SIMILARITY MEASURES

As said above, conceptual space theories define similarity as a measure that is inversely proportional to distance (i.e., between objects and/or the prototypes of concepts), that is usually determined according to a Minkowski metric. Let me recall the expression for the distance (in a generic Minkowski metric) between two objects (and/or prototypes of concepts) *a* and *b* located within an *n*-dimensional space:

$$d(a,b) = \left(\sum_{i=1}^{n} w_i \left| f_i^{[a]} - f_i^{[b]} \right|^p \right)^{1/p}$$

where $f_i^{[o]}$ (or $f_i^{[C]}$) represents the value of the *i*-th dimension of the object *o* (or concept *C*); w_i represents the weight assigned to the contribution of the *i*-th dimension; and the value of the parameter *p* determines the kind of metric (e.g., if $p=1$ the metric is city-block or Manhattan; if $p=2$ the metric is Euclidean).

The expression above applies to the standard/ordinary Minkowski distance. However, those distances might be weighted differently according to various criteria. For instance, the weight could be a function of the number of examples (i.e., instances or concrete cases) on which a particular concept is based. In such a case, the distance-of-comparison $d_C(o,P_C)$ between an object *o* and a concept *C* –represented within the conceptual hyperspace by the prototype P_C– may be expressed under a multiplicatively weighted scheme[13] as follows[14]:

$$d_C(o,P_C) = u_C \, d(o,P_C)$$

where u_C represents the weight assigned to the distances from the prototype of *C* (i.e., P_C) to any other point of the conceptual space[15]. In section 4.3.5 below, I will show the implications of non-standard weighting for the convexity requirement in Gärdenfors' conceptual spaces.

VORONOI DIAGRAMS (OR DIRICHLET TESSELLATIONS)

In a similarity-based space theory of concepts, the categorization of an object *o* under a particular concept is the result of a mental process that (i) evaluates the distances from *o* to the prototypes of all the relevant concepts within the considered context, and (ii) classifies *o* under the closest concept –that is, under the concept *C* whose prototype P_C is the most similar to *o*–. Once a given similarity measure is adopted, a similarity-based

[13] For a detailed review of weighting approaches that are distinct from the multiplicative one, see Okabe *et al.* (1992, pp. 119-134).

[14] *Ordinary distances* use to be called simply *distances* –or *standard distances*–, and that is what I will do in my thesis. In addition, I will use the terms *weighted distance* and *non-standard distance* indistinctly.

[15] Under this kind of weighting, similarity in a conceptual space resembles to the force of gravity in a gravitational field, where gravity –or similarity– is not only inversely proportional to distances, but also directly proportional to the mass of the attracting body –or to the size (measured in terms of the number of examples) of the considered concept, respectively–.

conceptual space can be characterized by means of Voronoi diagrams, inasmuch as concepts may be conceived as the cells resulting from a Voronoi tessellation of the conceptual hyperspace (see Fig. 4.3), whose input are the prototypes of the relevant concepts.

Voronoi diagrams date back to mid-nineteenth century[16], in particular to the theory of quadratic forms and its interpretation by means of Voronoi partitions (Gauss 1840), that led Dirichlet (1850) to prove the unique reducibility of quadratic forms. Then, at the beginning of the twentieth century, Voronoi (1908) generalized Dirichlet's result to higher dimensions. In order to honor their work, this kind of construct receives the name of Voronoi diagrams or Dirichlet tessellations[17].

A Voronoi diagram is a partition of an *n*-dimensional space into regions, based on the distances between each point and the points belonging to a particular subset *G* of that *n*-dimensional space. The points belonging to *G* are commonly called *seeds* or *generators* and, in a prototype-based approach, those points are the *prototypes* of concepts[18]. The general idea is that for each generator g_i there is a region constituted by those points nearest to g_i than to any other seed belonging to *G*. The points equidistant from their two closest generators will constitute the boundaries of regions.

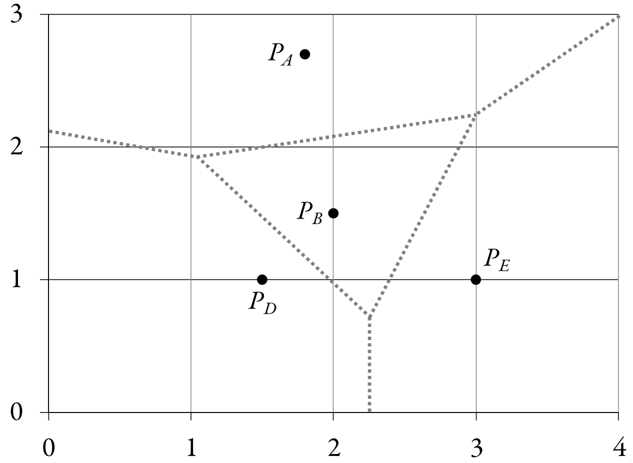

Fig. 4.3. Boundaries of the conceptual regions resulting from the tessellation of a Euclidean conceptual hyperspace, by means of a maximization process following the principles of the prototype theory of concepts. The final prototypes P_j are represented by the four black dots with coordinates (1.5,1), (1.8,2.7), (2,1.5) and (3,1). The boundaries of the conceptual regions are represented by means of grey dotted lines.

[16] Nonetheless, to the extent that Voronoi cells (i.e., the regions resulting from a Voronoi tessellation) are commonly identified with convex polytopes, Voronoi diagrams might be traced back at least to Descartes (1644), so are known to mathematicians since that time (Gruber 2007).

[17] For a review of the Voronoi diagrams history, algorithms and applications, see Aurenhammer (1991).

[18] This set of prototypes is not static, but dynamic, given that their locations are updated after the subject's exposure to new examples of those categories.

For instance, Fig. 4.3 shows a Voronoi tessellation of the conceptual space with a set G constituted by the points P_j representing the prototypes of four concepts A, B, D and E (i.e., the points P_j are the generators g_i of the Voronoi partition). If the standard Euclidean metric is taken –where p=2, and both concepts and dimensions are equally weighted–, then the boundaries of regions will be determined by the bisectors of the segments connecting each pair of prototypes, as displayed in Fig. 4.3.

4.1.4 Strengths of the geometric conceptual space theory

As claimed in previous chapters, even with violations of the metric axioms (i.e., minimality, symmetry and triangle inequality)[19], the geometric approach to similarity models has remarkable advantages, such as its ability to identify the underlying dimensions in a particular data set, the production and handling of spaces with reduced dimensionality –what is quite convenient for coding, memory and processing–, and its application in categorizations and inferences[20,21]. On this basis, it may be said that the main strengths of a geometric similarity-based space theory of concepts are the following:

(1) *Same explanatory framework for objects, concepts and properties*: the first reason for the geometric approach is its unifying character, since it can coherently integrate in a unique theory the principal elements (i.e., objects, concepts and properties) present in any discussion about the nature and structure of concepts. The idea is that the geometric view locates objects and concepts in the same representational framework, namely the conceptual hyperspace, in which properties play a natural structural role as a proper element of the model –more concretely, as the constitutive factors of that conceptual hyperspace[22]–.

[19] Recall that, as was explained in section 3.3.4, the defenders of the geometrical approach to similarity resorted to context in order to account for why the metric axioms were sometimes violated. The idea was that in similarity judgments (i.e., in the evaluation of similarity between two objects and / or concepts), context-dependent factors may intervene, which would convert those similarity judgments into directional ones, what would explain the potential violation of the axioms. And although a characterization of context is advisable, such a question is well beyond the scope of my thesis.

Thus, in geometric models like the one described in section 4.1.3 (which used a similarity measure based on multiplicatively-weighted distances-of-comparison $d_C(o,P_C) = u_C\, d(o,P_C)$), similarity judgments have directional character –that is, they are sensitive to the direction of comparison–, and this would explain the violation of both the axiom of symmetry, and the inequality triangle. Anyhow, this is not the only available explanation for the geometrical view. Indeed, another possible account is that some parameters of the model (e.g. the relevant dimensions, their weighting, the kind of metric, etc.) depend on the term of the comparison that acts as referent, which would make the similarity judgment into directional and, consequently, explain the violation of the axioms.

[20] Let me remind that another benefit of the similarity-based geometrical view was that, given that it is based on the prototype theory, it can easily explain typicality effects (see chapter 2).

[21] And, even though it would be useful to analyze how a geometric model might explain –or be compatible with– the empirical evidence concerning structural and transformational similarities (see sections 3.5 and 3.6), such study lies beyond the scope of my present work.

[22] Hence, by contrast to the connectionist approach –where properties are not explicitly represented–, the geometric models do represent properties in an explicit way.

(2) *Empiricist explanation of concept acquisition*: additionally, the geometric view allows to clarify how concepts could be acquired from empirical evidence about particular objects and/or events, even though the issue of the nature and origins of the most basic dimensions still remains problematic[23]. As said above in section 4.1.1, concepts –or, more concretely, the prototypes of concepts– could result from an optimization process that, taking as starting point the locations of objects within a dimensional space, maximized the similarities –or, alternatively, minimized the distances– between the points that represent objects belonging to the same category. Therefore, an approach like this explains how concepts can be acquired –or transformed– from experience about objects in the external world.

(3) *Easy account of memory, learning, categorization and inference*: a similarity-based geometric model of concepts also has great explanatory power, since it provides easy accounts for significant cognitive phenomena[24]: (a) *Categorization* is possibly the most direct application of the geometric view since, as explained in section 4.1.1, an object is said to belong to a given concept if and only if the values of that object produce an *n*-tuple that lies inside the region associated with the considered concept[25,26]. (b) *Memory* –in the sense of storage– is more feasible when conceived as operating on low-dimensionality spaces, like those produced in these models (through techniques such as multidimensional scaling or factor analysis)[27, 28]. (c) *Learning* beyond concept acquisition –that is, in the sense of generalization / specification of previously learned concepts–, which can be conceived as the union/separation, respectively, of the regions associated to other already existing concepts. (d) *Inference*, both *inductive* –which occurs when a new concept is produced from the information about particular objects, as described in the prior point–, and *deductive*[29] –which comes about when the representation associated

[23] Chapter 6 will be precisely devoted to this question, namely to the issue of how the formation of primitive concepts (i.e., the most basic elements of a conceptual system) can happen. This issue is directly related with the need of reliable ontogenetic models that explain how those primitive conceptual constituents may be acquired.

[24] Additionally, Voronoi tessellations can easily explain other phenomena, such as conceptual change (Gärdenfors and Holmqvist 1994) and conceptual vagueness (Douven *et al.* 2013; Douven 2016).

[25] Consequently, it could be said that categorizations are a straight consequence of the main thesis of this type of models.

[26] Additionally, I will assume that in this kind of approach *reference determination* can be reduced to *categorization* tasks. In this case my point is that, under a contextualist perspective like the one in this thesis, concepts always depend on context, so there is no normativity of meaning that may be appealed in order to fix reference. Thence, according to this view what determines the extension of a concept in our cognitive system would be the execution of a particular categorization algorithm.

[27] The lower the memory resources required to store the same information, the higher the global amount of information that can be stored. Similarly, the lower the number of variables required to characterize a particular concept/property, the faster the access to that concept/property.

[28] Nevertheless, in order that categorization and memory are computationally efficient, concepts must be conceived as prototypes, and not as conceptual regions. For more on this, see section 5.2.2.

[29] For a study of how the conceptual space theory can explain deductive inferences see Hautamäki (1992).

to a particular object/concept is completely included in the region associated to another concept–.

(4) *Cognitive economy*: additionally, an approach based on Voronoi partitions supports –in a cognitively efficient way– many of the aforementioned psychological processes like memory, concept learning, and categorization[30]: (a) *Memory* –in the sense of storage– is more efficient since only the locations of the prototypes have to be stored. (b) *Concept learning* can be carried out from a small number of particular examples of the considered category. (c) In *categorization* tasks, the only distances evaluated are those between the considered object and the prototypes associated with the relevant categories. Consequently, it is fair to say that a Voronoi-based articulation of the geometric model of similarity is highly efficient from a cognitive point of view.

(5) *Ability to operate autonomously on the basis of elementary processes*: finally, another virtue of the geometric approach is that it can be articulated in a way such that it is based on very simple processes, that may even be independent of any other conceptual information or background knowledge stored by the mind. The idea here is that the two main basic processes which will be proposed in the following chapters (i.e., processes of dimensional reduction and data clustering) can plainly work on raw –or slightly processed– perceptual data. The strong point of an empiricist model like this is that higher level cognitive capabilities (such as pattern recognition, learning of new concepts and properties, etc.) can emerge from such a simple characterization of the processes which underlie our conceptual system.

4.2. Gärdenfors' conceptual spaces

Gärdenfors' approach to the similarity-based space theories of concepts –or, more succinctly, Gärdenfors' *conceptual spaces*– is undoubtedly the most important and influential geometrical framework in the current debate. His proposal was initially characterized in his book *Conceptual Spaces* (2000), and fourteen years later clarified and extended in *The Geometry of Meaning* (2014).

Nevertheless, although Gärdenfors' approach is the best-known version of the conceptual space theory, there are relevant differences between his proposal and the general framework –as we will see in the following subsections–. In the first place, according to Gärdenfors (natural) *properties* are convex regions of a given domain (Criterion P), and are typically associated to the meaning of adjectives. This contrasts with the general framework, where properties are either the constitutive factors of the conceptual hyperspace, or (convex or non-convex) regions in a particular conceptual space. Secondly, (natural) *concepts* are said to be bundles of properties –or, alternatively, sets of convex regions– in a number of domains, together with the salience weights of those domains and information concerning how their regions are correlated (Criterion C), typically representing the meaning of nouns. Finally, although in Gärdenfors' view the notion of

[30] Concept storage and categorization –i.e., points (a) and (c)– will be addressed in chapter 5 of this thesis, while concept acquisition –i.e., point (b)– is the topic of chapter 6.

domain –defined as a set of integral dimensions that are separable from all other dimensions– is critical, such a notion plays no role in the general approach to the conceptual space theory.

The aim of this section is to provide a detailed introduction of Gärdenfors' conceptual spaces, as well as of his main assumptions, thesis and applications.

4.2.1 Motivation

In order to understand Gärdenfors' reasons for developing his theory of conceptual spaces, it is convenient to begin with his conception of meaning within the cognitive semantics framework. According to Gärdenfors (1996; 1999), if semantics is defined as the relation between linguistic expressions and mental structures, then conceptual spaces constitute an appropriate framework for those cognitive structures, since they can provide a suitable ontology for such a semantics (Gärdenfors 2000, p. 159). This approach to cognitive semantics is based on the following six tenets –which are openly admitted by Gärdenfors (1996; 1999; 2000)–:

(I) *Cognitivist conception of meaning*: the first and main principle is that meanings are conceptual structures in cognitive models –and not truth conditions in possible worlds–. And, given that cognitive semantics is conceived as a mapping between language and mind, on Gärdenfors' view no reference to reality is needed to determine the meaning of a linguistic expression[31]. Additionally, truth –as a relation between the mind and the world– is considered by Gärdenfors to be subsequent to –and, consequently, independent of– meaning.

(II) *Meanings are perceptually grounded*: since the conceptual structures in the mind depend partially on perception, then meaning is tied to body-experience. Based on this, Gärdenfors argues that the constitutive dimensions of his conceptual spaces can be grounded on perception. (But, as said in footnote 31 of this chapter, this second tenet is not completely compatible with a strong interpretation of the first one.)

(III) *Cognitive structures have a geometrical character*: in Gärdenfors' view, meaning is constituted by conceptual elements whose nature is essentially geometrical, and it is not identified with symbols related by a system of rules.

(IV) *Cognitive structures are not essentially propositional, but image-schematic*[32]: in this case the geometrical character of the conceptual structures proposed by

[31] Nonetheless, this latter claim might not be the case. Indeed, this point in tenet (I) is hardly compatible with tenet (II), since while the first holds that meaning is independent of reality, the second maintains that meaning is a function of perception. Of course, Gärdenfors could argue that once the physical experience has been integrated in the cognitive system, such information may be thought to be part of the conceptual structure. However, that seems to be a quite weak –and vague– argument. Anyhow, this issue could be explained in a better –and more concrete– way: since in Gärdenfors' approach meaning is a function of the dimensions associated to a certain domain, and given that the domain may depend on context, then meaning will be a function of that context and, in consequence, dependent on reality.

[32] Tenets (IV) and (V) are –possibly– the most controversial ones.

Gärdenfors allows to explain metaphors –or at least some of them– as geometric transformations of those mental structures, in line with the *spatialization of form hypothesis* proposed by Lakoff (1987, p. 283), according to which concepts can be understood as spatial image schemas modulated by metaphorical mappings[33].

(V) *Semantics has (temporal) primacy over syntax*: on Gärdenfors' view semantics has temporal precedence over syntax, since perceptual representations are conceived as previous to the development and intervention of language. Indeed, he even accepts that the cognitive structures constrain the syntax employed to express the semantic elements[34].

(VI) *Conceptual spaces are compatible with the prototype theory of concepts*: Gärdenfors' conceptual spaces provide a natural explanation for the typicality effects identified by Rosch (1973; 1975; 1983) in many concepts, which could not be done from the standpoint of the classical theories of concepts (see chapter 2), given that: (a) the centres –or centroids– of the regions associated with each concept in a conceptual space may be identified with the most representative members in a prototype theory; and (b) prototypicality can be characterized as a measure inversely related with the distance from an object to any of those centres.

4.2.2 Gärdenfors' geometric approach

With regard to the main elements present in Gärdenfors' geometrical approach, some of them are identical to their corresponding notions in the general framework (e.g. *distances* and *regions*), while others either have relevant differences (e.g. *dimensions*, *properties* and *concepts*), or are specific of Gärdenfors' proposal (e.g. *domains*). The present section will be focused on these two latter groups.

Dimensions

According to Gärdenfors, conceptual spaces are built up of –or constituted by– (quality) *dimensions*, on which objects, properties and concepts will be represented. Thence, the major role of dimensions is to represent the different qualities of objects (Gärdenfors 2000, p. 6). Possible examples of dimensions include *time*, *temperature*, *pitch*, *size* and *weight*.

Those dimensions will be the coordinate axes that constitute the underlying conceptual hyperspace. For instance, the dimension of *weight* would be a one-dimensional

[33] In particular, Lakoff distinguishes six kinds of conceptual structures which are identified with the following schemas: (i) *categories* would correspond with CONTAINER-schemas; (ii) *hierarchical structures* with PART-WHOLE and UP-DOWN-schemas; (iii) *relational structures* with LINK-schemas; (iv) *radial structures* with CENTER-PERIPHERY-schemas; (v) *foreground-background structures* with FRONT-BACK-schemas; and (vi) *linear quantity scales* with UP-DOWN and LINEAR ORDER-schemas.

[34] On my view, this is the most controversial thesis since, although semantics has to be phylogenetically prior to syntax (i.e., before having the means to express something, it is necessary to have something to be expressed), it not so clear that the same remains valid once syntax is developed. In fact, it might be the case that the same mutual influence recognized by many between semantics and pragmatics, also took place in the case of the relationship between syntax and semantics.

linear structure that takes values from the positive real line. By contrast, although the dimension of *time* also has linear structure, it takes values from the whole real line, where "present" corresponds to the point zero, and "past" and "future" to the negative and positive real lines, respectively[35].

Therefore, in Gärdenfors' view properties should not be confused with dimensions, and this is a relevant departure from the general conceptual space theory. On the one hand, sometimes both notions collapse in Gärdenfors' conceptual spaces. That occurs in the case of one-dimensional domains (e.g. domains of *time*, *temperature*, *pitch* and *weight*), whose properties are also one-dimensional. But, on the other, in most of the cases domains will be multi-dimensional (e.g. the domain of *size*, which is constituted by the three spatial dimensions, namely *height*, *width* and *depth*; or the domain of *color*, which is conceived as constituted by three dimensions, i.e. *hue*, *saturation* and *brightness*) by virtue of which, in Gärdenfors' conceptual spaces, properties cannot be identified with the basic dimensions of their associated domains[36].

Finally, Gärdenfors holds that these (quality) dimensions should be seen, not in scientific terms (i.e., not as elements of a scientific theory), but as cognitive entities –or constructs– through which a plausible interpretation of the world is provided.

Domains

Gärdenfors utilizes a quite specific notion of domain. According to him, dimensions are involved in stable groups (i.e., a dimension like *hue* does not appear alone but together with the dimensions of *saturation* and *brightness*). Those stable sets of related dimensions are called domains. On this basis, Gärdenfors defines a *domain* as "a set of integral dimensions that are separable from all other dimensions"[37] (Gärdenfors 2000, p. 26). Thus, for instance, the *color* domain would be constituted by the dimensions of *hue*, *saturation* and *brightness* (see Fig. 4.4); and the domain of *space* by the dimensions of *height*, *width* and *depth*[38]. As said above, domains can be one-dimensional, in which case they would be constituted by only one dimension.

[35] And with regard to the issue of dimensional structure, Gärdenfors is also disposed to accept coordinates systems different from the Cartesian –in particular, *polar coordinates* for the representation of locative prepositions (Gärdenfors 2014, pp. 205-217)–.

[36] Obviously, it could be argued that maybe properties based on multi-dimensional domains might be reduced to one-dimensional domains. For example, even though the *size* domain is constituted by three dimensions, once such a domain is built up and sizes are determined, those size values can be rearranged over only one dimension. In this case Gärdenfors uses to reply that such a one-dimensional rearrangement is not possible for the case of *color*. Nonetheless, and independently of whether that rearrangement is possible –either directly, or through dummy variables–, perhaps Gärdenfors' excessive focus on the *color* domain led him to turn an exception (i.e., the domain of *color*) into a rule (i.e., his distinction between the notions of *dimension* and *property*).

[37] Gärdenfors' motivation for splitting cognitive structures into domains seems to be the possibility that properties in one domain (e.g. the *color* of an object) may be characterized independently of the properties belonging to other domains (e.g. the *size* or *weight* of that object) (Gärdenfors 2014, p. 22).

[38] Nevertheless, not every domain considered by Gärdenfors is as clear-cut as the *color* and *space* domains. For example, in the case of the *taste* domain –which, on Gärdenfors' view, would be constituted by the

An important distinction for this definition of domain is the one between separable and integral dimensions. At this point Gärdenfors relies on the work in cognitive psychology, which claims that integral dimensions are those processed in a holistic and unanalyzable way, where the assignation of a value to a particular dimension requires a value to be given to the others (Garner 1974; Maddox 1992; Melara 1992). By exclusion, if dimensions are not integral, then they are separable[39]. For instance, in the *color* domain the dimensions of *saturation* and *hue* are integral because perception does not seem to give a *saturation* value without also assigning a value to the dimension of *hue*; in the *sound* domain, the dimensions of *pitch* and *loudness* are said to be integral because the interaction between them is so strong that subjects do not usually distinguish them. By contrast, the dimensions of *shape* and *color* are called separable, since their stimuli use to be analyzable; and the same happens for the dimensions of *size* and *lightness* (Handel and Imai 1972).

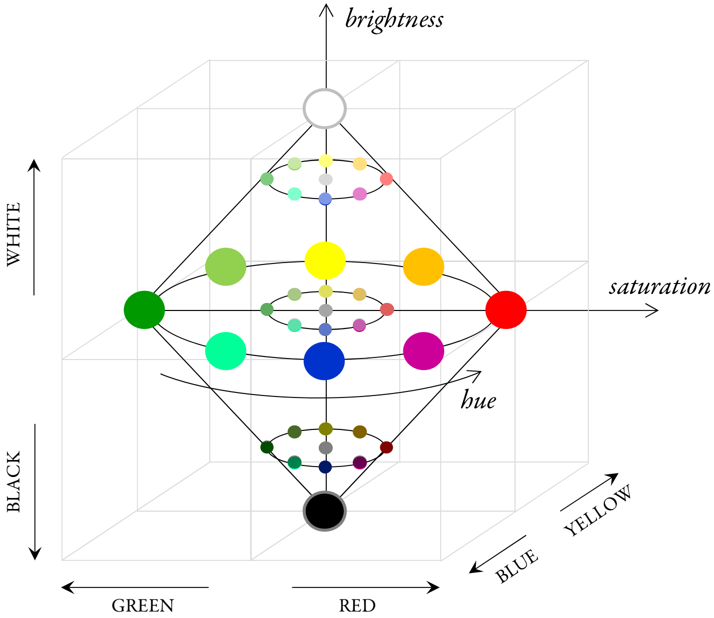

Fig. 4.4. Representation of the domain of color and its constitutive dimensions (i.e., *hue*, *saturation* and *brightness*) (adapted from Churchland 2005, p. 536): (a) *hue* has circular structure, so it is represented by an angular coordinate; (b) *saturation* –or *intensity*– has linear structure with only positive values, and it is represented by the radial coordinate (i.e., by the distance from the vertical –or longitudinal– axis); and (c) *brightness* has linear structure with two endpoints, and it is represented by the vertical axis.

dimensions of *salt*, *bitter*, *sweet*, *sour* and maybe another fifth dimension (Gärdenfors 2014, p. 21)–, it is not clear which topological structure might tie together those five dimensions, nor the fact that they all are integral (i.e., non-separable) dimensions.

[39] Many times it is argued that separable dimensions fit better with a city-block metric, while a Euclidean metric is preferable with integral dimensions (Hyman and Well 1968).

Properties and Concepts

Next, based on the notion of domain, Gärdenfors summarizes his definitions for (natural[40]) *properties* and *concepts* in what he calls Criteria P and C[41]:

> Criterion P: A *natural property* is a convex region of a domain in a conceptual space. (Gärdenfors 2000, p. 71)

> Criterion C: A *natural concept* is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated. (Gärdenfors 2000, p. 105)

Therefore, according to Gärdenfors *properties* and *concepts* are two different families of the same conceptual class (i.e., the class whose elements are represented by regions in a mental space). So, while the notion of *property* is used to denote information in only one domain, the notion of *concept* might be applied to information based on one or more domains[42, 43]. Consequently, under this view properties are a special case of concepts –in particular, properties are concepts determined by regions based on only one domain–.

For instance, in Gärdenfors' approach color terms such as "red" and "blue" express natural properties –or, in other words, RED and BLUE are natural properties–, because they can be represented by means of the three constitutive dimensions of the *color* domain. In fact, a programmatic thesis in Gärdenfors' work is that the properties conveyed by simple words (i.e., adjectives) in natural languages are natural properties:

> *Single-domain thesis for adjectives*: The meaning of an adjective can be represented as a convex region in a single domain. (Gärdenfors 2014, p. 136)

Clearly, this is a very strong and controversial thesis, since it entails that the meaning of any adjective may be characterized by means of a set of dimensions –constitutive of one particular domain– whose values can be set independently of the values assigned to the

[40] Although these two definitions are for *natural* properties and concepts, as a matter of fact Gärdenfors applies them almost universally. Indeed, he does not distinguish between *natural* and *non-natural* properties/concepts, except in order to discriminate artificial non-convex properties/concepts, such as those associated with Goodman's (1955) term *grue* (i.e., green before a given date and blue after that date).

[41] This definition of property was originally introduced ten years earlier by Gärdenfors (1990), and both of these two definitions may be found –in nearly the same terms– in his most recent book (Gärdenfors 2014, pp. 24 and 124).

[42] Obviously, the distinction between properties and concepts is only relevant if the notion of domain is accepted, by virtue of which properties may be defined in terms of the dimensions of only one domain. Nonetheless, and given that such a notion plays no role in my proposal, here I will depart from Gärdenfors' position and follow the traditional view –according to which properties and concepts belong to the same class–.

[43] This contrasts with the equivalent role played by properties and concepts in other areas (e.g. in first order logic and λ-calculus, where both of them –properties and concepts– are represented by means of homogeneous predicates).

dimensions of any other domain. Nonetheless, under a contextualist perspective like the one adopted in my doctoral thesis, the *single-domain thesis for adjectives* simply cannot be the case, given that if concepts –and properties– always depend on context, then it is not possible to identify a closed set of dimensions, belonging to only one domain, which allow to determine every possible context. For instance, observe the strong differences between the adjective "tall" –a clear case of single-domain adjective–, and the adjective "handsome" –a good candidate for (context-dependent) multiple-domain adjective–.

With regard to his definition of *concept*, in Gärdenfors' view concepts are not mere bundles of properties (i.e., convex regions in a number of domains), since information about the weights of –and the correlations between– the distinct domains is also considered. For instance, the concept APPLE could be represented by the set of convex regions associated to following properties (Gärdenfors 2000, p. 102; Fiorini, Gärdenfors and Abel 2014, p. 132):

— RED, YELLOW and GREEN in the *color* domain;
— EPICYCLOID in the *shape* domain;
— SWEET and SOUR in the *taste* domain;
— SMOOTH in the *texture* domain;
— SUGARS, FIBER, VITAMINS, etc. in the *nutrition* domain; and
— SEED-STRUCTURE, FLESHINESS, PEEL-TYPE, etc. in the *fruit* domain.

The concept APPLE could also involve a strong –and positive– correlation between the domains of *taste* and *nutrition*, in particular between the properties SWEET and SUGARS (i.e., sugar-content) (Gärdenfors 2014, p. 25). Additionally, the distinct domains might be differently weighted, so that *color* and *shape* were the two most salient domains, and *taste* and *nutrition* the two least salient ones.

Lastly, another specific element in Gärdenfors' proposal is the requirement of convexity on the geometric structure of the regions representing properties –and, consequently, concepts–. However, since section 4.2.5 below is especially devoted to the description of Gärdenfors' convexity constraint, and section 4.3 contains a detailed discussion of the difficulties associated with such a requirement, at this point I will say no more about the convexity issue.

4.2.3 Notion of concept

In the previous section I have described Gärdenfors' original definition of concept, as formulated in *Conceptual Spaces* (2000) –which merely applied to object categories–. Nevertheless, in his latest book *The Geometry of Meaning* (2014) Gärdenfors extends his notion of concept from object categories to action and event categories. Since my work will be mainly focused on the first group –namely, object categories–, it seems advisable to see how that notion of concept (i.e., concepts as object categories) has varied through those last fifteen years.

First, let us recall Gärdenfors' definition of concept by means of CRITERION C, which only applied to concepts in the form of object categories:

> Criterion C: A *natural concept* is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated. (Gärdenfors 2000, p. 105)

From then on, Gärdenfors' notion of concept has suffered a significant change, as it is evident when we read his most recent definition of it:

> An *object category* is determined by
> (i) a set of relevant domains (may be expanded over time)
> (ii) a set of convex regions in these domains (in some cases, the region can be the entire domain)
> (iii) prominence weights of the domains (dependent on context)
> (iv) information about how the regions in different domains are correlated
> (v) information about meronomic (part-whole) relations. (Gärdenfors 2014, p. 124)

In regard to the latter definition, it must be said that criteria (i) to (iv) are essentially contained in his original proposal, and allow an interpretation of concepts as convex regions –criteria (ii) and (iv)–, whose dimensions are differently weighted depending on domain –criterion (iii)–, and where criterion (i) constitutes a specification of the conceptual hyperspace that contains the region associated to the considered concept.

Finally, in accordance with the general framework, Gärdenfors' concepts are the result of splitting the similarity space into (convex) regions constituted by sets of points which represent the objects that exhibit the properties characteristic of those regions. And, again, these (convex) regions are identified with concepts.

4.2.4 Extensions and applications of the conceptual space framework

Besides, in his most recent work, Gärdenfors and collaborators have tried to extend the original framework from the case of properties and concepts –or, alternatively, from adjectives and nouns– to the representation of changes, actions and events; through these, to the semantics of verbs, adverbs and prepositions; and ultimately to apply it to the case of human communication (Gärdenfors and Warglien 2012; Warglien *et al.* 2012; Warglien and Gärdenfors 2013; Gärdenfors 2014). Very briefly sketched, Gärdenfors takes as a starting point the thesis that *verbs* typically express dynamic properties of objects which –as parts of *events*– use to involve *actions* that may be described in terms of *forces* –commonly exerted by agents–. By virtue of this, his proposal for verbs consists of a holistic model of actions, forces and events, characterized by means of convex regions in conceptual spaces. In such a framework, *verbs* express changes in properties (that is, movements in the representation of objects –or concepts– within the conceptual space), and are represented by convex regions of vectors.

Furthermore, some supporters of Gärdenfors' conceptual spaces have already tried to extend this framework to areas beyond theories of concepts. For example, some of them have proposed Gärdenfors' theory in order to model the phenomena of meaning negotiation and of "meeting of minds" (Warglien and Gärdenfors 2013, 2015; Gärdenfors 2014), both of them within the debate about human communication. Others have suggested the possibility of applying the conceptual spaces framework to the field of sensory cognition, in particular to the areas of the sensory perception of vision, smell, taste and touch (Paradis 2015), and also to the case of music perception (Chella 2015). Lastly, research has been undertaken to model changes in scientific frameworks, when their elements are described in terms of spatial structures like those proposed in the conceptual spaces theory (Gärdenfors and Zenker 2013; Zenker and Gärdenfors 2015a)[44].

4.2.5 The convexity requirement

As is evident from previous sections in this chapter, the requirement for the convexity of regions runs through all the conceptual space theory defended by Gärdenfors. Not only properties and concepts –or object categories–, but also the semantics of verbs, adverbs and prepositions (Gärdenfors 2014) are conceived and represented within his theory by convex regions.

The *convexity requirement* can be thought of as a generalized definition of the conception of natural kind as a qualitatively spherical region (Quine 1969, p. 119). However, when characterizing the geometrical structure of natural properties and concepts, three possible criteria may be distinguished in order to constrain the geometry of a region (see Fig. 4.5):

- *Connectedness constraint*: it must be possible to reach every point in the region from every other point by following a continuous path consisting only of points belonging to the region.
- *Star-shapedness constraint* (with respect to a point P): for every point x in the region, all the points between[45] x and P must belong to that same region.
- *Convexity constraint*: the region must satisfy the star-shapedness constraint with respect to all the points in the region, that is, for every two points x and y in the region all the points between x and y must also belong to that same region.

The strength of these three criteria increases in order (i.e., every star-shaped region is a connected region, and –trivially– every convex region is a star-shaped region).

Gärdenfors opted for the strongest of those criteria (i.e., the convexity constraint), mainly due to these three reasons[46]: (a) mutual dependence with the prototype theory; (b) cognitive economy; and (c) perceptual foundation.

[44] For a more exhaustive list of the possible extensions and applications of Gärdenfors' conceptual space theory, see Zenker and Gärdenfors (2015b).

[45] Although the second and third constraints do use the notion of *betweenness*, whose axiomatic definition may be found in Gärdenfors (2000, p. 15), I will not address that issue in my thesis.

[46] Since section 4.3.2 below contains an exhaustive critical discussion of Gärdenfors' arguments in favor of the convexity of conceptual regions, at this point I merely enumerate those main reasons. Additionally, in section 4.3.2 I will hold that none of Gärdenfors claims is a compelling reason to accept convexity as a compulsory requirement for the geometry of conceptual regions.

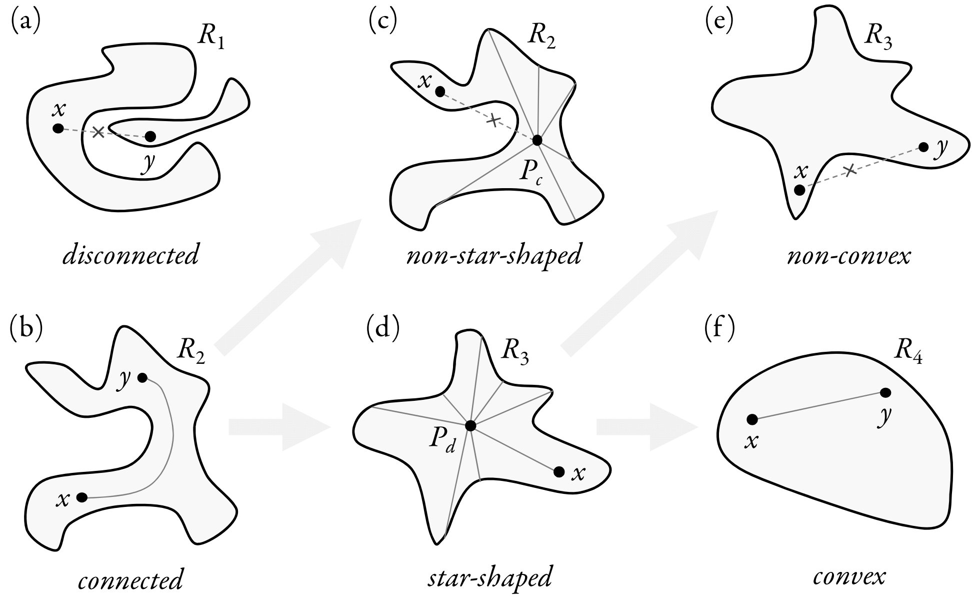

Fig. 4.5. Representation of the three different criteria for the geometry of conceptual regions (i.e., *connectedness*, *star-shapedness* and *convexity*). Paths containing exclusively points belonging to the considered regions are represented by solid lines, while paths containing also points outside the regions are represented by *dashed lines*. The problematic points not belonging to the regions are represented by means of crosses (×): (a) Representation of a *disconnected* region R_1 where the point x is not reachable from the point y through a continuous path of points belonging to R_1. (b) Representation of a *connected* region R_2 where every point x in R_2 is reachable from every other point y in R_2, following a continuous path of points belonging to R_2. (c) Representation of a *connected* but *non-star-shaped* region R_2 –the same as in graph (b)–, where there is no point with respect to which R_2 satisfies the star-shapedness constraint; for instance, P_c cannot be that point because between P_c and x there are points not belonging to R_2). (d) Representation of a *connected* and *star-shaped* region R_3, where there is a point P_d in R_3 such that, for every other point x in R_3, all the points between P_d and x also belong to R_3. (e) Representation of a *star-shaped* but *non-convex* region R_3 –the same as in graph (d)–, where there are points x and y in R_3 such that not all the points between them also belong to R_3. (f) Representation of a *star-shaped* and *convex* region R_4, where for every two points x and y in R_4, all the points between x and y also belong to R_4.

4.3. The role of convexity in conceptual spaces

As seen in section 4.2, one of the main theses involved in Gärdenfors' conceptual spaces is that the regions associated with properties, concepts, events, verbs, etc. are convex. Indeed, the convexity constraint is the crucial condition which distinguishes Gärdenfors' conceptual spaces from any other geometric articulation of the prototype theory of concepts. Or, in other words, given that Gärdenfors' proposal adds nothing to the (geometric) similarity-based theories of concepts –excepting the convexity demand on the shape of conceptual regions–; without such a condition, his theory of conceptual spaces is equivalent to a contextualist similarity-based approach to concepts –like that of the prototype theory– characterized by means of a geometric model[47]. Therefore, if it could be proved that convexity is an unnecessary condition on the shape of conceptual regions, then we could speak of conceptual spaces with no reference to convexity (in a purely general geometric characterization of the prototype theory of concepts), just as I will do in chapters 5 and 6 of this thesis.

The aim of this section is to show that the convexity constraint on the geometry of conceptual regions is: (a) *unnecessary*, from a theoretical perspective; and (b) *problematic*, when some particular applications of the theory are considered (Hernández-Conde 2017b). First, I show that all the arguments provided by Gärdenfors in favor of the convexity of regions rest on controversial assumptions. Next, I claim that his argument in support of a Euclidean metric, based on the integral character of conceptual dimensions, is weak, and under a non-Euclidean metric the structure of regions can be non-convex. Furthermore, even if the metric is Euclidean, the convexity constraint may be not satisfied if concepts were differentially weighted. Lastly, I hold that Gärdenfors' convexity constraint is brought into question when considering some particular applications of his conceptual spaces since: (i) some of the allegedly convex properties of concepts are not convex; (ii) the conceptual regions resulting from the combination of convex properties can be non-convex; (iii) convex regions may co-vary in non-convex ways; and (iv) his definition of verbs is incompatible with a definition of properties in terms of convex regions.

By virtue of all this, and since the acceptance of non-convex conceptual regions has little or no influence on the cognitive efficiency and explanatory power of the theory, I will claim that mandatory character of the convexity criterion in the conceptual space theory must be reconsidered. In such a case (i.e., if the convexity criterion were abandoned), the conceptual spaces theory would be no more than a contextualist geometric articulation of the prototype theory of concepts, and that is exactly the kind of position assumed by me through the rest of my thesis.

4.3.1 Voronoi tessellations in Gärdenfors' theory

My aim in this section is to show that Gärdenfors' theory implicitly accepts the thesis that the shapes and boundaries of conceptual regions[48] are produced by means of a Voronoi partition of the conceptual hyperspace, whose inputs are the prototypes of the relevant concepts. (From here on I will call it THESIS V[49].) This is important because, as

[47] Both of them –i.e., Gärdenfors' conceptual spaces and a (contextualist) geometric articulation of the prototype theory – explain the same phenomena (e.g., concept acquisition and change, categorization, inference, conceptual vagueness, etc.) based on the same input data, and they do it in the same way.

[48] Here I use the expression "conceptual regions" in a general sense, in order to refer to the regions associated to concepts, but also to properties, verb meanings, etc.

[49] THESIS V is a thesis commonly assumed by most advocates of the dimensional approaches to the prototype theory of concepts and, more particularly, assumed by me in chapters 5 and 6.

said in section 4.3.2 below, THESIS V is compatible both with convex and with non-convex conceptual regions.

Firstly, THESIS V is tacitly present when explaining how significant elements of the theory work. For example, Gärdenfors' account of categorization, conceptual learning, conceptual change, communication and language, and conceptual vagueness are all of them based on THESIS V (Gärdenfors 2014, ch. 2 and appendix). And, since no alternative explanation is considered for all those processes and phenomena, different from the one based on Voronoi tessellations, it may be claimed that: although Gärdenfors never openly expresses his commitment to THESIS V, this thesis is tacitly accepted by his theory of conceptual spaces.

Nonetheless, the regions produced by a Voronoi tessellation are only convex under very specific assumptions, namely if the metric is Euclidean and if distances are not differently weighted for each particular concept (see sections 4.3.4 y 4.3.5). Thus, Gärdenfors could try to defend the convexity constraint through two different strategies: either arguing directly in favor of the convexity of the conceptual regions; and/or arguing for the assumptions which guarantee that the regions resulting from a Voronoi partition are convex. In the next subsections I will try to show that: [I] Gärdenfors' direct arguments in favor of convexity are not conclusive. [II] The assumption of a Euclidean metric is not guaranteed. [III] It is not implausible that the distances to the prototypes of distinct concepts are differently weighted.

4.3.2 Gärdenfors' arguments for the convexity constraint

In his work, Gärdenfors does not provide any definitive argument in favor of the convexity constraint, but he offers a series of reasons that suggest a high degree of plausibility for it. My aim in this section is to show that none of those arguments (i.e., mutual dependence with the prototype theory, cognitive economy, and perceptual foundation) compels us to accept convexity as a compulsory requirement for the geometry of conceptual regions[50].

[50] Although in Gärdenfors (2000) he distinguished between *properties* –defined as convex regions– and *concepts* –defined as sets of convex regions–, it is not possible to find there an explicit assertion about the convexity or non-convexity of *concepts*. However, things change in his latter works, where we find statements like "as proposed in Gärdenfors (2000), concepts can be modeled as convex regions of a conceptual space" (Warglien and Gärdenfors 2013, p. 2171), or "the convexity of concepts is also crucial for ensuring the effectiveness of communication" (Gärdenfors 2014, p. 26). Nonetheless, it is not clear that in these last quotes Gärdenfors is referring with the term *concept* the same as in Gärdenfors (2000), because in Gärdenfors (2014) the notion of *object category* began to play the role played by the notion of *concept* in Gärdenfors (2000). Withal, in Gärdenfors (2014), he does not explicitly assert that object categories are represented by convex regions.

Notwithstanding this, in these recent works (Warglien and Gärdenfors 2013; Gärdenfors 2014) he tries to explain human communication via a "meeting of minds", using a fix-point argument. Such an argument requires the convexity of concepts: "the concepts in the minds of communicating individuals are modeled as convex regions in conceptual spaces (...) If concepts are convex, it will in general be possible for interactors to agree on joint meaning even if they start out from different representational spaces" (Warglien and Gärdenfors 2013, p. 2165). But, this argument is expected to apply at least to noun phrases (*ib.*, p. 2170) and, thus, to object categories –as the APPLE concept–, which should also be

Mutual dependence with the prototype theory

As seen in section 4.2.1, one of the six tenets that Gärdenfors considers to be embodied by his cognitive approach to semantics is that concepts show prototypical effects which cannot be explained from the standpoint of the classical approach to concepts. Indeed, one of the major advantages of Gärdenfors' proposal was that his theory provides a natural explanation of prototypical effects for many concepts[51]. Let us see why.

As said above in this chapter, Gärdenfors' conceptual spaces are a particular type of the dimensional approach to the prototype theory of concepts. Due to this, an object *o* is categorized or not under a particular concept *C* in terms of the similarity between the object *o* and the concept *C*. Additionally, that similarity is determined by virtue of their shared properties. Besides, prototypes are acquired through a process of maximization of similarities between the evaluated objects and the tentative prototype, and as a consequence the prototype of a given category will be its most typical member –whether real, or not[52]–. Thence, Gärdenfors' approach is able to explain typicality effects like those identified by Rosch and Mervis (1975) because, the more prototypical[53] a member of a category is (a) the more attributes it shares with other members of that category, and (b) the fewer features it has in common with the members of other categories.

With regard to this first claim, Gärdenfors defends the idea that those who adopt the prototype theory should expect a representation of concepts and properties as convex regions; and contrariwise, that if concepts are characterized as convex regions, then prototypical effects must be expected (Gärdenfors 2000, pp. 86-87; 2014, pp. 26-27).

It is my view that this is the main argument offered by Gärdenfors in support of the convex geometry of regions. However, neither of these two assertions constitutes a reason in favor of the convexity requirement, given that both of them could also be applied to a star-shaped region resulting from a Voronoi partition, as I now explain:

[A] *If properties and concepts were defined as star-shaped regions* (produced by a Voronoi tessellation of the conceptual space) *then prototypical effects would also be expected*: in this case the typicality of an object with regard to a given category is also a function of the distance between the point representing that object and the prototype of its category.

[B] *The only thing that should be expected by a consistent prototype theorist is the star-shapedness of conceptual regions*[54]: a prototype theorist should expect that if an ob-

represented by convex regions. On my view, this proves that Gärdenfors is committed to the convexity of both *properties* and *concepts*.

[51] See section 2.2.1.

[52] That is, with or without a real instance –or exemplar– of it.

[53] I recall that, under this view, prototypicality can be characterized as a measure that is inversely proportional to the distances between prototypes and/or objects.

[54] This is valid for the case of a standard metric (i.e., non-weighted metric), where all prototypes are equally-weighted. For a discussion of the case of a non-standard (Euclidean) metric, see section 4.3.5. (Other examples of partitions of a conceptual space by means of non-standard metrics may be found in Figs. 5.4 and 5.5 of chapter 5.)

> ject belongs to a certain category, then all the objects with the same proportional distances from the prototype but more similar to it –that is, all the objects between the object under consideration and the prototype–, should also belong to that category. This is exactly what happens under the star-shapedness constraint[55]. Therefore, the assumption of the prototype theory does not entail the convexity of regions, but only makes desirable their star-shapedness.

Thus, Gärdenfors' alleged mutual dependence between the prototype theory and the convexity of regions –in a conceptual space approach articulated by means of Voronoi partitions–, also takes place between the theory of prototypes and the star-shapedness of regions. Consequently, such a relationship cannot be a crucial reason in favor of the convexity constraint[56].

Cognitive economy

When Gärdenfors originally defined properties in terms of convex regions, he mainly based his decision on the argument provided by Shepard (1987, p. 1319). Shepard argued that evolution would have led to consequential regions (within our psychological space) in a way such that the boundaries of those regions were not oddly shaped. Next Gärdenfors (2000, p. 70) claimed that such evolutionary preference could be supported by a principle of *cognitive economy* in terms of memory, learning and processing.

Nonetheless, the cognitive economy argument depends on the assumption that the handling of convex sets of points requires less memory, learning and processing resources than the handling of regions with capricious forms. In this case, Gärdenfors' argument can be structured as follows:

[55] The argument which connects the prototype theory –articulated by means of Voronoi tessellations– with the star-shapedness of regions can be summed up as follows:

> Premise 1: If an object o belongs to a concept C (characterized by a prototype P), this entails that the ball $B(o, oP)$ –centered at o and with radius oP– does not contain any other prototype distinct from P. [Premise 1 is equivalent to the thesis that concepts are the result of a Voronoi tessellation (Thesis V).]
>
> Premise 2: Minkowski metric with $p \geq 1$ and non-weighted prototypes.
>
> Conclusion: Conceptual regions are star-shaped.

The general idea is that for every object a between o and P, it is possible to prove that the ball $B(a, aP)$ is included within $B(o, oP)$. Therefore, P is the nearest prototype to a; that is, the object a also belongs to C and, in consequence, conceptual regions are star-shaped. (For the specific details of this proof see Lemma 5 in Lee (1980, p. 608).)

[56] Withal, this should not be seen as a defense of a mandatory star-shapedness requirement, since I am disposed to accept a different weighting of concepts / prototypes –and in such a case the conceptual regions might be non-star-shaped (as shown in Fig. 4.7b)–; in contrast with others who propose to substitute the convexity criterion by the star-shapedness one. (By instance, Bechberger and Kühnberger (2017) put forward to substitute convex regions / sets by star-shaped regions / sets, on the basis of the fact that the convexity criterion prevents a geometric representation of the correlations between dimensions; a geometric representation that may be carried out by means of star-shaped sets.)

(i) Properties and concepts are determined by convex regions within a conceptual hyperspace. (Criteria P and C)

(ii) Those convex regions can be the result of a Voronoi tessellation starting with a set of prototypes[57]. (Thesis V)

(iii) Voronoi tessellations provide cognitively efficient explanations of psychological processes such as memory, concept learning and categorization (see section 4.1.4).

(iv) Therefore, the handling of convex regions can explain cognitive efficiency in all those tasks and processes.

The problem is that, as shown in sections 4.3.3 and 4.3.4 below, the conceptual regions could be non-convex and yet compatible with a Voronoi tessellation. That is, if Voronoi partitions are the source of the cognitive economy in the argument above, but they can produce both convex and non-convex conceptual regions; then, convexity cannot be obtained by abduction from the premise of evolutionary preference for cognitive economy, but only a characterization of concepts based on Voronoi tessellations.

All in all, cognitive economy is common to any conceptual space theory which assumes that concepts are represented by the regions resulting from a Voronoi partition, independently of the geometrical structure –either convex or non-convex– of those conceptual regions. In consequence, cognitive efficiency cannot be a crucial reason to support the convexity requirement.

Perceptual foundation

Gärdenfors also alleges that many perceptually grounded domains, such as *color*, *taste*, *vowels*, etc., are convex based on evidence in favor of the convexity of the regions associated with numerous typical properties of all those domains (Fairbanks and Grubb 1961; Sivik and Taft 1994). Indeed, the *color* domain seems to be his preferred example. However, the problem of the *color* domain –when used as evidence of convexity– is that it is entirely associated with sensory dimensions, and there is no guarantee that things work in the same way in non-perceptual domains. This last point is explicitly recognized by Gärdenfors (2014, p. 137) when he acknowledges that such evidence –mainly associated with the *color* domain– does not provide automatic support for the convexity constraint in other domains.

* * *

Thence, the problem is that none of these reasons is compelling enough to accept convexity as a mandatory constraint on the geometry of regions. In fact, all of them are highly questionable: (a) the argument based on the mutual dependence with the prototype theory could also apply to any (non-convex) star-shaped region resulting from a Voronoi tessellation; (b) the cognitive economy one depends on the controversial assumption that handling a Euclidean metric is computationally less demanding than handling other sorts of metric; and (c) the claim of perceptual foundation relies on the presumption that

[57] In order that the resulting regions may be convex, the Voronoi tessellation will have to meet some conditions (i.e., Euclidean metric and non-weighted prototypes).

perceptual and conceptual (i.e., non-perceptual) domains share the same geometric structure.

All in all, on the one hand, under the assumption of Thesis V neither the mutual dependence with the prototype theory of concepts, nor the cognitive economy argument, is a definitive reason in favor of convexity. This is so because those arguments are equally valid for any conceptual division produced by a Voronoi tessellation, independently that their shapes are convex or not. On the other hand, the other argument for convexity (i.e., perceptual foundation) does not urge to engage with the thesis that concepts have to be convex.

4.3.3 Integral dimensions, Euclidean metric, and convexity

As stated above, Gärdenfors' conceptual spaces rest on the assumption that regions are convex and, if the underlying metric were the standard Euclidean metric (that is, if distances were Euclidean and non-weighted), then the convexity of regions would be guaranteed (Gärdenfors 2000, p. 88; Okabe *et al.* 1992, p. 57). By virtue of this, the theory requires that the metric underlying our psychological space is Euclidean. On this occasion, the main argument in favor of a Euclidean metric is that, for the case of integral dimensions[58], the Euclidean metric is more suitable than the city-block metric. (By contrast, the latter would be more appropriate for the case of separable dimensions.) And, because Gärdenfors' definitions of property and concept are for domains constituted by sets of integral dimensions, it may be alleged that the conceptual spaces underlying them function with a Euclidean metric and, consequently, that their associated regions are convex.

However, this argument presents several problems, mainly due to the adduced mutual dependence between integral dimensions and the (standard) Euclidean metric:

— First, Gärdenfors alleges a sort of co-implication between integral domains and the Euclidean metric: "If the Euclidean metric fits the data best, the dimensions are classified as integral; (...) when the dimensions are integral, the dissimilarity is determined by both dimensions taken together, which motivates a Euclidean metric" (Gärdenfors 2000, p. 25). The first implication is true, since if the metric is not city-block, the dimensions are non-separable (i.e., they are integral). Nonetheless, the second conditional is false, inasmuch as the non-separability of dimensions does not necessarily imply that the metric is Euclidean[59].

[58] Remember that, as said in section 4.2.2, dimensions are integral if they are those processed in an unanalyzable way (i.e., in a way such that assigning a value to any of them requires giving a value to the others).

[59] On the one hand, the city-block metric is accepted when dimensions are separable because, in such a case, the dimensions are the most meaningful element: they contribute independently to the total distance, and must remain invariant (i.e., unrotated) to keep the same conceptual space structure (Garner 1974, p. 119). On the other hand, the Euclidean metric is proposed when dimensions are not analyzable because, in this case, distance (which remains the same for all rotations of axes) is the most relevant element.

Thus, the Euclidean character of the metric structure cannot be based on the integral character of domains. Gärdenfors assumes that dimensions are integral, but that is not sufficient to guarantee that the metric is Euclidean. Therefore, the Euclidean structure of a metric space needs empirical evidence independent from the one associated to the non-separability of its constitutive dimensions.

— Secondly, and even assuming, as Gärdenfors does, that the co-implication between integral dimensions and the Euclidean metric were the case, the empirical evidence referred to in favor of the integral or separable character of a particular set of dimensions is tied to perceptual domains[60], such as *color*, *sound*, *size*, *shape*, etc. (Garner 1974; Maddox 1992; Melara 1992). All that work faces a threefold difficulty, when taken as evidence in favor of Gärdenfors' theses, as I now explain:

[A] The experiments were developed over a small number of perceptual domains, so accepting them as evidence of the geometry of conceptual spaces requires the assumption that the behavior of the metric structure is the same across all perceptual and conceptual domains[61]. That is, it is necessary to assume that such behavior extends, not only from the studied perceptual domains to all other perceptual domains, but also to all conceptual domains –in general not related to any of the perceptual domains studied–, which might not be the case.

[B] This kind of work is used to contrast the Euclidean metric with the city-block metric, and shows that the former fits integral sets of dimensions better, while the latter provides a better fit when dimensions are separable. In the case in hand, however, the problem is that both metrics provide *good* fits, but not *perfect* fits. For instance, Handel and Imai (1972, p. 110) showed that the optimal parameter *p* for integral dimensions in a general Minkowski metric is 1.7, which means that the best metric is neither the Euclidean nor the city-block one, but something between these two[62].

Therefore, what can be derived from Handel and Imai's work is not that the (standard) Euclidean metric is warranted for integral dimensions, but only that the expected metric for integral domains will be closer to the (standard) Euclidean metric than to the (standard) city-block metric.

However, it is possible that (i) although the dimensions did not contribute independently to the value of distances (and, therefore, distance were the most meaningful element); (ii) it happened that, despite (i), the conceptual space structure were not irrelevant, and distances were not invariant under rotations of dimensions. In that case, dimensions will be non-separable –by virtue of (i)– but, at the same time, they will not be secondary –by virtue of (ii)–. Therefore, the non-separability of dimensions does not imply the Euclidean structure of the underlying conceptual space, whose metric could be non-Euclidean (for instance, with *p* equal to 1.7 or 3) and still able to explain the non-separability of domains.

[60] As happened with the perceptual foundation argument for the convexity constraint (see section 4.3.2).

[61] This behavior of the metric structure can be summed up as follows: separable dimensions are better characterized by a city-block metric, while the Euclidean metric is the best for integral dimensions.

[62] And, for a parameter p=1.7 the conceptual regions are still non-convex. (See section 4.1.3 for the meaning of the p parameter within the standard Minkowski metric; and see Fig. 4.6c for a chart which shows that conceptual regions are not convex for a parameter p equal to 1.7.)

[C] It is doubtful that the integral character of a group of dimensions is something innate or immutable, not even in the case of perceptual dimensions. Indeed, there is evidence that such integrality may be altered by high-level cognitive processes; and even that it can be the remains of a developmental trend toward more and more differentiated –i.e., separated– dimensions. By instance, it has been proved that –by means of training– *saturation* can be differentiated from *brightness* (Burns and Shepp 1998; Goldstone 1994b); and also that dimensions easily isolated by adults –e.g., *brightness* and *size*– are handled as joined together by 4-year old children (Kemler and Smith 1978; Smith and Kemler 1978)[63].

In summary, all this evidence appears to be controversial; both that supporting the integral character of conceptual dimensions, and that which allegedly backs up the relationship between the integral character of dimensions and the Euclidean metric. The result is that, in both cases, the underlying metric could be non-Euclidean and, in consequence, conceptual regions could be non-convex.

4.3.4 Conceptual spaces under a non-Euclidean metric

Nonetheless, merely attending to the basic requirements of a similarity space theory of concepts, it can be seen that the convexity constraint is unnecessary. Indeed, nothing in the general conception of these theories demands a Euclidean metric, and under a non-Euclidean metric the conceptual regions resulting from a Voronoi tessellation can be non-convex. Anyhow, a constant throughout all of Gärdenfors' work is that he explicitly adopts a Euclidean metric which –apparently– guarantees the convexity of the conceptual regions. The problem is that if the conceptual space metric is non-Euclidean, then regions may be non-convex, so the aim of this section is to describe what the consequences would be under the assumption of a non-Euclidean metric.

As introduced in section 4.1.3 above, the formula for the (standard) distance, given a generic Minkowski metric, between two objects –and/or prototypes of concepts– *a* and *b* located within an *n*-dimensional space, is given by the expression:

$$d(a,b)=\left(\sum_{i=1}^{n} w_i \left| f_i^{[a]} - f_i^{[b]} \right|^p \right)^{1/p}$$

where the value of *p* determines the type of distance (e.g. p=1, Manhattan; p=2, Euclidean), and it may take any positive real value –not only integers– greater or equal to 1. The boundaries of the regions will then depend on the chosen metric, and so will do their convex or non-convex character (as illustrated by the graphs in Fig. 4.6).

As is evident from Fig. 4.6, only the (standard) Euclidean metric satisfies the convexity requirement[64], while all the other metrics produce conceptual regions that are more or

[63] For a review of the evidence in favor of the developmental trend in the direction of increasingly separable dimensions, see Goldstone and Steyvers (2001).

[64] For a formal demonstration of this, see Okabe *et al.* (1992, p. 57).

less non-convex. In consequence, if the metric of conceptual spaces is not Euclidean in a strong sense, then the convexity constraint on the shape of regions cannot be mandatory in a strong sense; contradicting what Gärdenfors' theory requires of them.

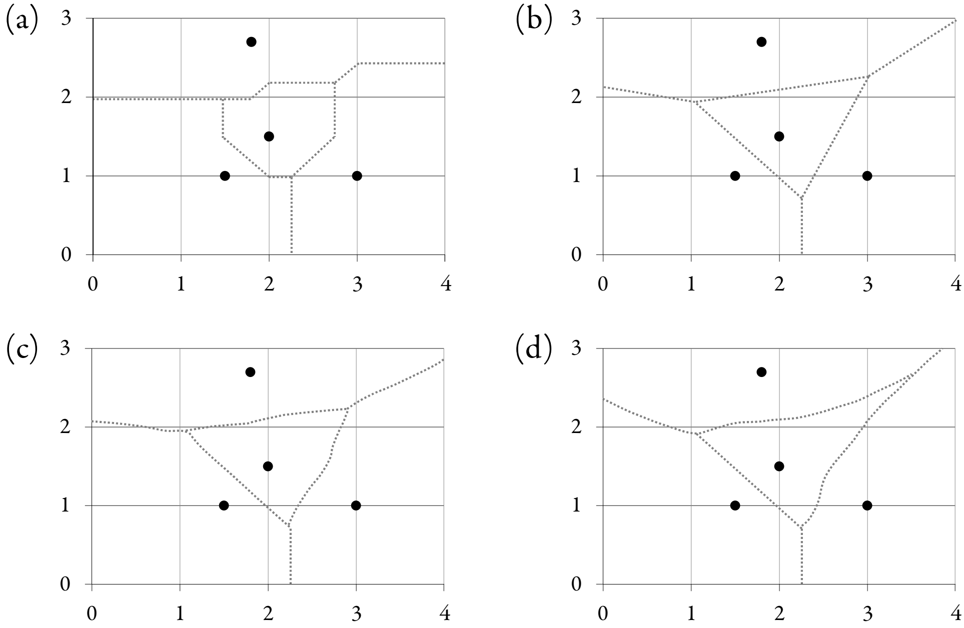

Fig. 4.6. Boundaries of the conceptual regions resulting from a Voronoi partition for four possible distinct metrics. The final prototypes are represented by the four black dots, whose coordinates are (1.5,1), (1.8,2.7), (2,1.5) and (3,1). The boundaries of the conceptual regions are drawn as dotted grey lines. (a) Boundaries for the city-block metric (parameter p=1). (b) Boundaries for the Euclidean metric (parameter p=2). (c) Boundaries for a conceptual space that fits the Euclidean metric better than the city-block one (parameter p=1.7) –as proven in Handel and Imai's (1972, p. 110) experiments–. (d) Boundaries for a higher-order Minkowski metric (parameter p=3).

4.3.5 Conceptual spaces under a weighted Euclidean metric

Nevertheless, even if the metric of conceptual spaces was Euclidean, it is possible that conceptual regions were not convex. Certainly, this could not be case, as just explained in the section above, under the standard Minkowski distance which, for the Euclidean case (parameter p=2), is defined as follows:

$$d(a,b)=\sqrt{\sum_{i=1}^{n}(f_i^{[a]}-f_i^{[b]})^2}$$

But let me first remind how, in Gärdenfors' theory, concepts are produced. To begin with, if a given concept is not innate, then it should have been learnt sometime in the past from a set of particular examples. Besides, it may be argued that the number of examples has an effect on how objects are categorized.

Consider, for instance, a subject *S* who had been exposed to hundreds of instances of the concept DOG, but only a few cases of the concept FOX. It could be thought that if that same subject were exposed to one new instance of FOX, different from all the foxes already encountered –e.g., an arctic fox–, but with a certain resemblance to the concept DOG already acquired, *S* might categorize the new instance under the concept DOG, and not under the concept FOX (see Fig. 4.7). The reason would be that, in a conceptual space theory of the mind, the weight u_{DOG} ascribed to the concept DOG can be different –and greater– than the weight u_{FOX} ascribed to the concept FOX[65].

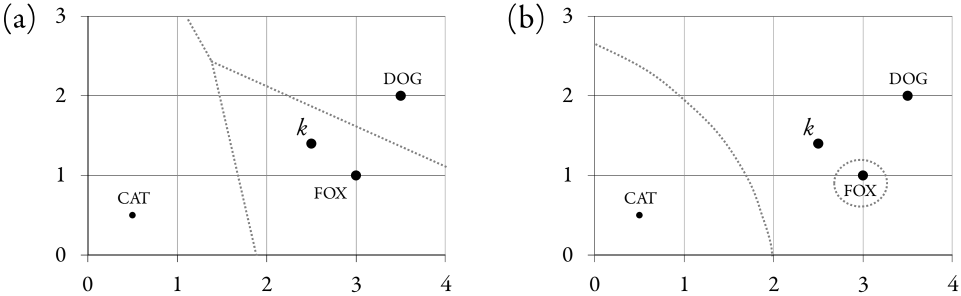

Fig. 4.7. Boundaries of the conceptual regions resulting from a Voronoi partition for different weightings of concepts. The three considered concepts are DOG, CAT and FOX, whose prototypes are represented by the black dots, with coordinates (3.5,2), (0.5,0.5) and (3,1). The boundaries of the conceptual regions are drawn as dotted grey lines. The point *k*, with coordinates (2.5,1.4), represents a case of arctic fox. (a) Boundaries for the standard Euclidean space, where the weights of all the prototypes are equal to 1. In this case the arctic fox represented by *k* is categorized under the concept FOX. (b) Boundaries for a non-standard prototype-weighted Euclidean space, where the distances-of-comparison for the concepts DOG, CAT and FOX are multiplicatively-weighted by 0.25, 0.4 and 1 respectively. (This could happen if the subject had been exposed (i) to a great number of instances of dogs, (ii) to a smaller number of cats, and (iii) only to a very few number of foxes.) In this second case, the arctic fox represented by *k* is classified under the concept DOG.

A phenomenon like this could occur even under a Euclidean metric –that is, even if the underlying conceptual space were Euclidean–, where base distances were calculated using the above formula for $d(a,b)$. If objects were categorized as just been described, the distance associated with each concept would be differently weighted depending on the number of examples on which that concept were based[66]. Besides, those differently weighted distances would correspond with the non-standard multiplicatively-weighted

[65] Obviously, the phenomenon of differently-weighted concepts is different from, and should not be confused with, the phenomenon of conceptual vagueness.

[66] A distinct weighting of concepts –by the number of instances of each concept– would be consistent with the frequency effects observed for the case of exemplars (Barsalou 1985; Nosofsky 1988b; Barsalou, Huttenlocher and Lamberts 1998).

distance introduced in section 4.1.3. In consequence, the formula for the distance-of-comparison $d_C(o,P_C)$ in categorizations of a particular object *o* with regard to a given concept *C* (represented by the prototype P_C) would be:

$$d_C(o,P_C) = u_C\, d(o,P_C)$$

Here, the value of the parameter u_C represents the weight associated with the concept *C*, which would be a function of the number of examples n_C on which such a concept is based. Indeed, the greater the number of examples n_C, the greater the relative similarity between *o* and P_C –or, in other words, the lower the distance-of-comparison $d_C(o,P_C)$–, and hence, the lower the weight of the distances u_C. The weight u_C could be, for example, a function ranging from two (if the number of examples is very small) to one (when that number is large enough), as given by the formula $u_C = 1 + 1/n_C$ (see Fig. 4.8).

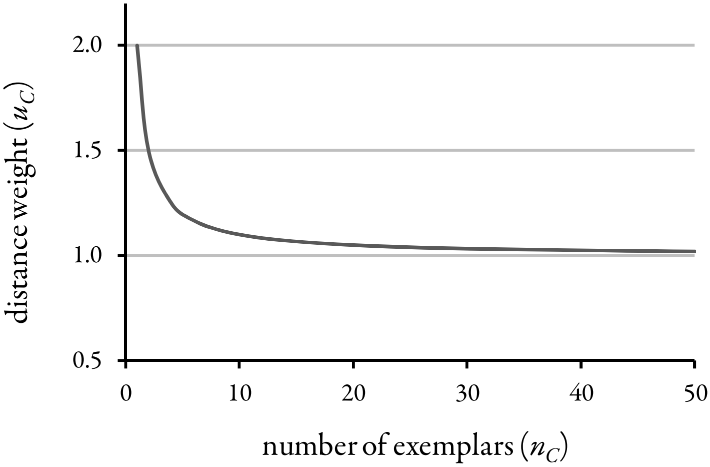

Fig. 4.8. Representation of the weight function $u_C = 1 + 1/n_C$, which could underlie a non-standard multiplicatively-weighted Euclidean space.

My point is that a conceptual space which functioned in this way would produce conceptual regions whose shapes are different from the ones produced by the standard Euclidean metric. Those shapes will be commonly non-convex, which contradicts the assumption regarding the convexity constraint. The graphs in Fig. 4.7 contrast the boundaries of convex regions in the standard Euclidean space, with the boundaries of non-convex regions in a prototype-weighted Euclidean space.

Thence, if concepts were differently weighted –depending on the sizes of their sets of examples– then, even within a Euclidean space, conceptual regions could be non-convex[67]. Of course, the foregoing requires empirical contrast through psychological

[67] For a summary of the properties of a weighted conceptual space, see Okabe *et al.* (1992, pp. 120-123). One of those properties is that the regions resulting from a multiplicatively-weighted Voronoi tessellation do not need to be convex –as shown in Fig. 4.7–, nor even connected; and that they can also con-

research, which will have to decide whether concepts are differently weighted or not. In spite of this, at least from a theoretical outlook, the size of the set of examples from which a certain concept is learnt could influence the reliability of such a concept. On my view, this possibility is significant by itself, beyond the fact that at present there exists or not conclusive empirical evidence about it.

Consequently, there are important reasons to think that not every concept has the same weight in the conceptual space structure. And if this were the case, then those distinct weights would lead to a non-standard Euclidean space, which would result in non-convex conceptual regions.

4.3.6 Problems of the convexity criterion in the working of Gärdenfors' theory

So far I have shown the following. [1] None of the arguments provided by Gärdenfors for the convexity constraint constitutes a compelling reason in favor of that requirement, given that all of them rest on controversial assumptions. [2] His argument for the integral character of conceptual dimensions –in support of a Euclidean metric– is weak, and under a non-Euclidean metric, the structure of regions will be non-convex. [3] Even if the metric were Euclidean the convexity constraint might be not satisfied –if, for example, distinct concepts were differently weighted in terms of the number of examples on which each of them is based–.

However it could be the case that, despite all of this, conceptual regions are in fact convex (as assumed by Gärdenfors). In this section, I show that Gärdenfors' convexity constraint is brought into question by his own characterization of conceptual spaces. On the one hand, I will show that in some cases the regions associated with the properties of a concept are not convex –either taken individually, or as the result of their combination in that concept–; while in other cases, the composition of convex regions associated with properties can lead to non-convex concepts –depending on how the properties co-vary over those concepts–. On the other hand, I will show that Gärdenfors' definition of properties in terms of convex regions is not compatible with his characterization of verb meanings as convex regions of vectors from one point to another.

On the convexity of properties and concepts

One of the most recent papers co-authored by Gärdenfors provides a detailed description, absent from previous work, of how his conceptual spaces work (Fiorini *et al.* 2014). In that paper, Gärdenfors and collaborators represent the inner structure of the APPLE concept by the product space resulting from the properties in those quality domains that form such a conceptual space (as shown in Fig. 4.9).

tain holes. Additionally, according to this kind of approach, the region associated with a particular concept *C* will be convex if and only if the weights of all its adjacent regions are greater than u_C (in Fig. 4.7 that is the case for the regions associated with the concepts CAT and FOX).

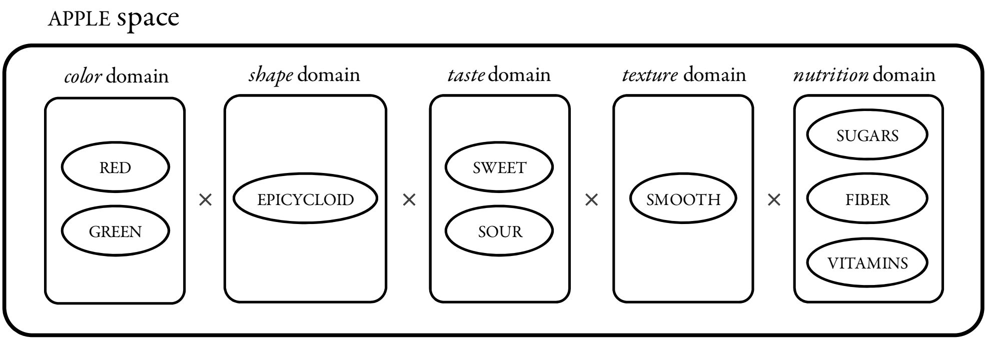

Fig. 4.9. Inner form of the APPLE *conceptual space*, as a product space of different (quality) properties. The APPLE space is represented by the bigger rounded rectangle. Quality properties (i.e., RED, GREEN, EPICYCLOID, etc.) are convex regions represented by the ellipses; for example, the property GREEN corresponds with a convex region in the *color* space –or *color* domain–. Quality domains (i.e., *color*, *shape*, *taste*, etc.) are represented by the smaller rounded rectangles. (Adapted from Fiorini *et al.* 2014, p. 132).

Difficulty 1 *Some of the properties are not convex.*

This difficulty could be summed up as follows. There are non-convex physical properties, and it is not easy to conceive a convex approach for the representation of some of those non-convex properties. The first point is largely uncontroversial, since the physical shape of many objects is not convex, as happens with the shape of an apple.

With regard to shape properties, Gärdenfors proposes different models for representing them, which are suitable for different kinds of shapes. Nevertheless, none of those models is proper for the characterization of general shapes and, in particular, for the convex representation of the shape of an apple, as it is shown in the next points:

(A) The approach followed in order to represent the concept RECTANGLE (Gärdenfors 2000, pp. 93-94; 2014, pp. 35-36) by means of the conditions satisfied by their quadruples of points in $\mathbb{R}^2$ can only be applied to very basic geometrical shapes (not including the epicycloid).

(B) The model proposed for the analysis of general shapes (Gärdenfors 2000, pp. 95-96; 2014, pp. 121-122), based on the work of Marr and Nishihara (1978), may be more or less applicable to the case of the shapes of animals[68], but not to the shapes of arbitrary objects.

❖ Therefore, neither model (A) nor model (B) allows us to represent the shape of the concept APPLE, distinguishing it from the shape of the concepts LEMON, PEAR or MELON. And, even though model (B) is useful for characterizing movements and actions, none of them is compelling as a model for general shapes.

[68] As a combination of cylinders –associated with their different parts–, together with information about how those cylinders are joined together.

(C) His approach to locative prepositions (Gärdenfors 2014, pp. 205-214) leads to an accurate formalization of the meaning of NEAR, FAR, INSIDE, OUTSIDE, BESIDE, etc. in terms of a polar coordinate system. In light of this, it seems that Gärdenfors' aim is to transfer how these prepositions are applied to shapes in the physical world, to the shapes of their associated conceptual spaces[69]. And the same can be said regarding his description of the meaning of BUMPY, as a structure in the physical space constituted by "an even (but continuous) distribution of values on the vertical dimension of a horizontally extended object" (Gärdenfors 2014, p. 246). However, a direct translation of shapes from the physical space to a convex representation within a conceptual hyperspace is only possible if the shape of the considered object is convex. The problem is that the shape of a huge number of objects is not convex, as happens with the EPICYCLOID[70] for the case of apples. An epicycloid is a plane curve generated by the path of a point on a smaller circle –with radius *r*– as that circle rolls around a larger fixed circle –with radius *nr*–. The epicycloid curve is determined by the following parametric equations:

$$x(\theta)=r(n+1)\cos\theta-r\cos[(n+1)\theta]$$
$$y(\theta)=r(n+1)\sin\theta-r\sin[(n+1)\theta]$$

Therefore, an apple shape could be associated with an epicycloid with a value of *n* equal to 1 or 2 (see Fig. 4.10). However, none of those 2D curves –and consequently, none of their associated 3D surfaces– is convex, since in both cases it is possible to find pairs of points within the regions they bound such that some points between them do not belong to the same region[71].

Finally, this problem extends from the apple example to many other object categories whose shapes are not convex. And, even though the fact that Gärdenfors is not able to provide a method for the convex representation of non-convex shapes –such as the shape of an apple– does not constitute a proof that no method exists; the lack of a method for the convex representation of non-convex shapes (e.g., the shape of apples) is evidence for

[69] The problem is that, for the convexity constraint to be met by the regions characterizing those prepositions, the convexity of the objects to which they apply is also necessary.

[70] Although Fiorini, Gärdenfors and Abel (2014) describe the apple's shape as a cycloid, in fact that shape corresponds to an epicycloid.

[71] Regarding this, it could be claimed that, if the radius *r* is the only parameter in those equations, then the representation of the shape of apples by means of the epicycloids parametric equations –proposed here by me– produces a class of shapes that is convex in the *shape* domain. In such a case, it would be true that that class of shapes is trivially convex, because if r_1 and r_2 are radii whose associated epicycloids *A* and *B* are shapes of apples, then every epicycloid *E* produced by a radius r_k, such that $r_1< r_k< r_2$, also represents the shape of an apple. Unfortunately, there exists another parameter in those equations, namely *n*, which determines the number of cusps of the epicycloid. Consequently, it might be objected that the class of shapes produced by the epicycloid equations does not form a convex region in the shape domain since, although the epicycloids with *n* equal to 1 or 2 are shapes of apples, there are values of *n* between 1 and 2 (i.e., 1.2, 1.25 or 1.5), whose associated epicycloids cannot be identified with the shape of an apple.

the difficulty of conceiving a natural way of representing a non-convex shape by means of convex conceptual regions.

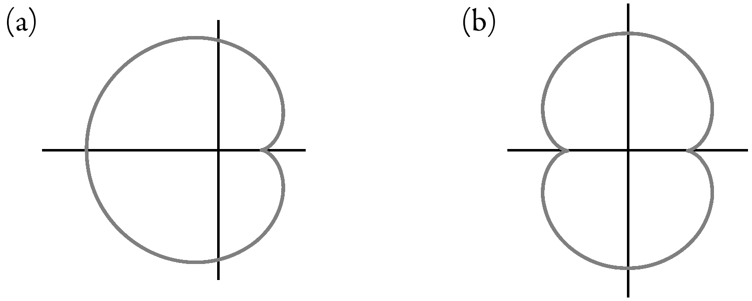

Fig. 4.10. Epicycloid curves representing the ideal two-dimensional (2D) *contour of an apple*. The apple's ideal shape will be the three-dimensional (3D) surface resulting from the rotation of any of these curves around the horizontal axis. (a) Epicycloid curve with ratio n=1. (b) Epicycloid curve with ratio n=2.

Difficulty 2 *The conceptual region resulting from the combination of convex properties (belonging to the same domain) can be non-convex.*

The characterization of the APPLE conceptual space (shown in Fig. 4.9) sheds a great deal of light on how conceptual spaces are supposed to work, especially regarding the following point: different properties in the same domain can be associated with the same concept (for example, the properties RED and GREEN in the *color* domain, or SWEET and SOUR in the *taste* domain, for the case of the APPLE concept).

Here the problem is that two properties from the same domain –for instance, RED and GREEN– cannot be composed into a product space. Let us recall that the product space R of a set of constitutive quality properties Q_1, Q_2, ..., Q_n, is equal to the set of objects[72] belonging simultaneously to Q_1, Q_2, ..., and Q_n. For instance, if the APPLE space were constituted only by the domains of *shape* and *texture*, then a particular object would be an apple if it were EPICYCLOID and SMOOTH. Or, from a logical point of view, for an object to be categorized as an apple, it is necessary –but not sufficient– that the following conjunction of properties is satisfied over those quality domains:

$$(\textit{shape} = \text{EPICYCLOID}) \wedge (\textit{texture} = \text{SMOOTH})$$

All this works if the properties considered belong to different domains. The problem is that when two –or more– properties belong to the very same domain, they cannot be

[72] It may be thought that the conceptual region R was equal to *the* (not *a*) set of objects belonging simultaneously to Q_1, Q_2, ..., and Q_n. However, that is not the approach followed by Gärdenfors (2014, p. 29), who accepts the possibility of non-rectangular conceptual spaces, which do not contain the whole sets of points belonging simultaneously to all their constitutive properties –as is the case of the graphs shown in Fig. 4.12–. That is the reason why the following logical conditions are necessary, but not sufficient.

composed into a product space, because in this case the product space would not include the desired set of objects. If we now included the *color* domain for the case of the APPLE concept, the resulting product space would have to satisfy the following condition:

$$(color_1 = \text{RED}) \wedge (color_2 = \text{GREEN}) \wedge (shape = \text{EPICYCLOID}) \wedge (texture = \text{SMOOTH})$$

This condition certainly includes all the red-and-green apples[73], but not the green (but not red) apples, nor the red (but not green) apples. This is so because the kind of composition required when two or more properties belong to the very same domain is not the *product* space, but the *addition* space, that is, the region resulting from the union of the regions associated with those properties. This kind of composition could be identified with a logical disjunction, so the condition associated with the APPLE space could be better expressed (with a unique *color* dimension) as follows:

$$[(color = \text{RED}) \vee (color = \text{GREEN})] \wedge (shape = \text{EPICYCLOID}) \wedge (texture = \text{SMOOTH})$$

The difficulty is that the conceptual space resulting from the addition of the RED and GREEN properties is not convex, because ORANGE establishes a discontinuity between them; as is obvious from their representation in the color spindle (see Fig. 4.11). In consequence, the resulting *color* space –associated with the APPLE concept– is not convex.

Lastly, an even clearer case is that associated with the SWAN conceptual space, which would be constituted (following Gärdenfors' approach) by the product space resulting from a set of properties in the quality domains of *color*, *shape*, etc. In this case, two different properties (i.e., BLACK and WHITE) are represented in the *color* domain. Those two properties would be represented by convex regions –in fact, points– within the *color* domain, but there is no path within the color spindle between them that only passes through points representing the colors of a swan. In this case, the combination of the BLACK and WHITE properties determines a disconnected region, so it cannot be convex (or even star-shaped)[74].

[73] At the expense of considering two different color dimensions (i.e., *color*$_1$ and *color*$_2$) as constitutive of the APPLE conceptual space, given that if both color dimensions were the same, the set of objects satisfying the condition would be void. Here I will not discuss the problems associated with such implications.

[74] Therefore, the SWAN conceptual space is a problem for any theory which attributes a mandatory character to the connectedness requirement for the geometry of conceptual regions. Obviously, this applies to any criterion stronger than the connectedness one –as is the case with the star-shapedness and convexity requirements–.

Nevertheless, I must recall that my work in this chapter should not be seen as an apology for a mandatory star-shapedness constraint, so cases like this are not a problem for my view. In fact, this sort of cases could be explained by means of the multi-prototype approach already mentioned in section 2.3.2 (Komatsu 1992), which is perfectly compatible with a (dimensional) similarity-based articulation of the prototype theory of concepts, like that considered in my work.

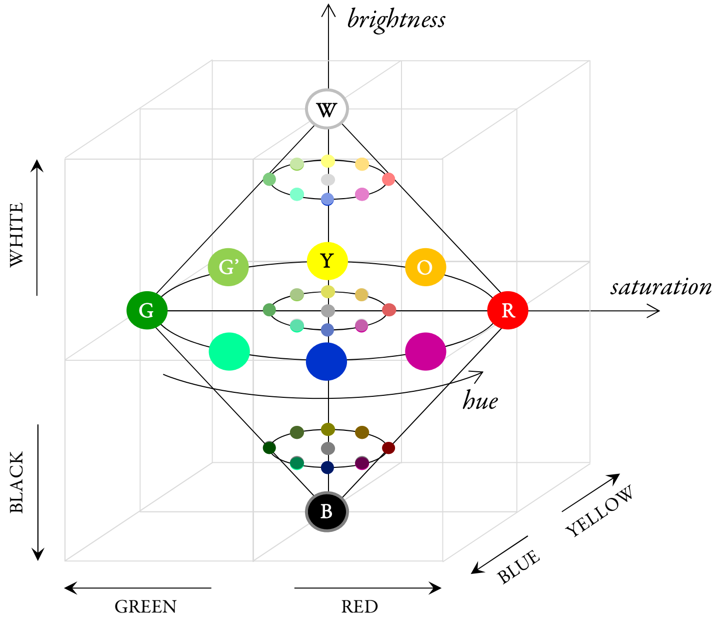

Fig. 4.11. Representation of the domain of color and its constitutive dimensions (i.e., *hue*, *saturation* and *brightness*) (adapted from Churchland 2005, p. 536). The relevant colors for the examples provided are denoted by their initials: B=black, W=white, G/G'=green, Y=yellow, O=orange, R=red.

Difficulty 3 *The conceptual region resulting from the covariation of convex regions can be non-convex.*

Even if a concept C is constituted by properties represented by convex regions (as Gärdenfors assumes), depending on how the properties of the instances –or particular cases– of C covary, the conceptual region R associated to C might be convex or not.

Let us now consider a concept C composed by two properties –or features– F_1 and F_2. If F_1 and F_2 were properties located in domains constituted by only one dimension (e.g. f_1 and f_2, respectively), then the region R representing C would be situated in a conceptual space composed by those two dimensions f_1 and f_2. That is the case of the MOUNTAIN concept, whose properties (i.e., WIDTH and HEIGHT)[75] are represented by one-dimensional domains (whose constitutive dimensions are *width* and *height*, respectively) (Adams and Raubal 2009, p. 258; quoted in Gärdenfors 2014, p. 29).

[75] Let me highlight the need of distinguishing between the *width* –or *height*– dimension, and the WIDTH –or HEIGHT– property. On the one hand, the *width* dimension is one of the coordinate axes which constitute the considered domain (namely, in this case the only coordinate axis of the homonymous one-dimensional domain of *width* –or *height*–). On the other, the WIDTH –or HEIGHT– property is a region within the conceptual space determined by the value –or range of values– taken by an object / concept in the *width* –or *height*– dimension.

Let me assume further that the properties F_1 and F_2 are represented by the regions –more specifically, intervals– Q_1 and Q_2 within the dimensions f_1 and f_2, respectively. In the particular case of the MOUNTAIN concept, if f_1 and f_2 represent the *width* and *height* of the mountain, respectively, those regions could be expressed in meters by the intervals $Q_1 = (1500, 13000)$ and $Q_2 = (1200, 8000)$.

Nonetheless, it could happen that the concept C were represented by a region R which did not cover completely the Cartesian product $Q_1 \times Q_2$ of its constitutive properties. In that case, there will exist pairs of points (q_1, q_2) which belong to $Q_1 \times Q_2$, but do not belong to R, that is, $\exists(q_1, q_2)((q_1 \in Q_1) \wedge (q_2 \in Q_2) \wedge ((q_1, q_2) \notin R))$. For instance, the MOUNTAIN conceptual region might not be rectangular, but triangular (see Fig. 4.12a):

> If a formation is very high, its width will not matter much; it will still be a mountain. However, a lower and very wide formation might not be called a mountain. Thus, the region in the product space that represents mountain has more or less a triangular shape. (Gärdenfors 2014, p. 29)

In such a case, an upward projection of the earth's surface whose WIDTH and HEIGHT were given by the pair (9000, 4000) –represented as k in Fig. 4.12a–, would belong to $Q_1 \times Q_2$, but would not be called a *mountain*.

Withal, if the conceptual space associated with MOUNTAIN is triangular, the convexity of the MOUNTAIN's conceptual region is guaranteed.

Nevertheless, if the region R (representing C) did not cover completely the Cartesian product $Q_1 \times Q_2$ of the constituent properties of C, it could happen that R is not convex. For instance, in Adams and Raubal's example the hypotenuse of the triangle which delimits the conceptual region of MOUNTAIN might not be a straight line, but a concave curve –as shown in Fig. 4.12b–. Therefore, depending on how the properties of a concept C covary, its conceptual region will be convex or not.

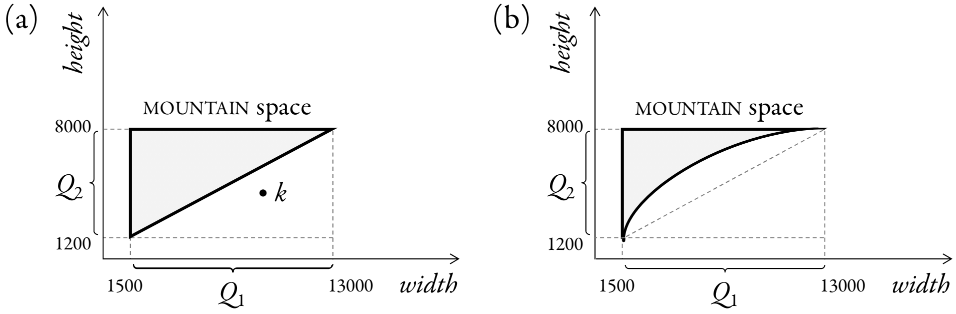

Fig. 4.12. Different ways in which the constitutive dimensions of the MOUNTAIN *concept* (i.e., *width* and *height*) *can covary*. (a) Covariation resulting in a triangular conceptual region. (b) Covariation resulting in a triangle with a concave curve as the hypotenuse.

ON THE COMPATIBILITY OF THE CONVEXITY OF PROPERTIES AND VERB MEANINGS

Ultimately, when Gärdenfors extends his conceptual space theory to the semantics of verbs, such an extended framework introduces a general difficulty (that could be described as structural). Such a problem is associated with his view on verb meanings, and closely related to his basic conception of property. In this case, the problem is that his characterization of verb meanings, in terms of vectors from one point to another, is not compatible with a definition of properties in terms of convex regions.

In his extended theory, Gärdenfors identifies *states* and *changes* with zero and non-zero vectors; and based on them he defines *events* as changes in the state of a patient –usually due to the action of an agent–. The problem is that, strictly speaking, *states* and *changes* cannot be identified with points and single vectors, respectively, if *properties* are not represented by points, but by (allegedly) convex regions. In virtue of this, *states* should be represented by regions; and therefore *changes* ought to be represented by sets of vectors from every point in the region associated with the initial property, to every point in the region associated with the final property[76].

The same can be said with regard to the result vectors associated with a given verb. Gärdenfors takes as starting point that verbs usually express changes in the properties of objects; that is, movements in the representation of objects/properties within the conceptual space. Based on this, *changes* are represented by means of vectors from the initial position of the object to its final position (i.e., from the position of the initial property to that of the final property). However, and given that a *state* is, in fact, not represented by a point, but by the region associated with the *property* described by that *state*, a verb may not be represented merely as a vector –or a mapping from one point to another–, but as a mapping from one region to another.

Thus, a verb should not be represented by a vector –or convex set of paths, with only one origin and one endpoint–, but by a set whose elements are convex sets of paths (each with a different origin and/or end).

Obviously, here I have not proved that those sets of vectors which should represent verbs cannot be convex. To my knowledge, it is hard to see in which sense it may be claimed that a set Z constituted by all the pairs of points (x,y) such that, $x \in X$ and $y \in Y$, where X and Y are convex sets, is convex. Notwithstanding this, the burden of proof lies on the side of Gärdenfors. If he wants (i) to provide a unifying framework for the semantics of verbs, nouns and adjectives (Warglien, Gärdenfors and Westera 2012), and (ii) to

[76] This is the same notion of *change* that Gärdenfors has in mind when he says that "[i]n general, a change of state is not represented by a specific vector. Instead, it can be represented by a category of changes of state. (...) If the start point is set as the origin, one can represent a category of change events as a region of end points. (...) For example, going 'upwards' in a two-dimensional space will correspond to a convex region of points located in a cone to the 'north' of the origin." (Warglien, Gärdenfors and Westera 2012, p. 162).

However, this is not the only possible way to generalize the notion of *change*. Another possibility would be that the knowing subject *S* knew neither the initial point nor the final point of the change, and that the initial point could not be set as the origin. This would happen if John said to *S*, "the leaves were yellowed by disease", but *S* does not know the tree whose leaves were referred to. In this case *S* knows neither the exact starting point (a kind of green) nor the exact end point (a kind of yellow) of the change expressed by "yellowed", so he will have to represent the properties associated to those initial and final states by regions (the ones associated to the GREEN and YELLOW colors). Therefore, the change expressed by "yellowed" will be represented as the set of vectors going from every point in the region representing GREEN to every point in the region representing YELLOW.

explain the semantics of verbs by means of changes of states (*ib.*), he has to prove that a set *Z* as the one just described is convex.

4.3.7 Conclusion

One of the main theses of Gärdenfors' (2000; 2014) conceptual space theory is the convexity constraint on the geometry of the conceptual regions associated with properties, concepts, actions, verbs, prepositions and adverbs. Nonetheless, in this section I have proved that such a requirement is problematic, both from a theoretical perspective, and with regard to some specific applications of the theory.

On the one hand, I have shown that none of Gärdenfors' arguments in favor of the convexity constraint compels us to accept it as a mandatory criterion for the geometry of regions. [1] With regard to his first argument, everything that can be said concerning the co-implication of the prototype theory and the convexity of regions, may also be said regarding the star-shapedness constraint on regions. [2] In relation to the cognitive economy argument, it depends on the controversial assumption that handling convex regions requires fewer computational resources than handling regions with arbitrary forms. [3] Finally, regarding the perceptual foundation argument, it relied on the hypothesis that perceptual and conceptual domains share the same geometrical structure, which might not be the case.

On the other hand, and with respect to the kind of metric underlying conceptual spaces, under the standard Euclidean metric assumed by Gärdenfors (i.e., Euclidean metric with non-differently weighted distances), the convexity of regions is indeed guaranteed. However, the question regarding the type of metric that underlies conceptual spaces is an empirical one; and all of the evidence provided by Gärdenfors in support of the standard Euclidean metric is controversial. Firstly, the Euclidean metric cannot be based on the integral character of domains, requiring empirical evidence independent from the one associated to the non-separability of dimensions. Secondly, the empirical evidence referred to in favor of the integral character of domains –and, in consequence, in favor of the Euclidean metric and the convexity of regions– comes from a very small number of perceptual domains; things might not work in exactly the same way in other perceptual and in non-perceptual domains. Thirdly, none of the works cited identify integral domains with the Euclidean metric perfectly, but rather with a metric that is more similar to the Euclidean than to the city-block; and such a kind of metric does not lead to convex conceptual regions. Considering all this, if the metric underlying conceptual spaces were standard, it could be that it was not Euclidean in a strong sense; and in such a case, it has been shown that the convexity constraint on regions is not valid.

Additionally, it has been proved that, even if the metric underlying conceptual spaces were Euclidean, regions could be non-convex if the distances-of-comparison in categorizations were differently weighted –depending, for example, on the number of examples on which each concept is based–. That is, convexity is guaranteed only under the standard Euclidean metric, but not under a weighted Euclidean metric. The problem is that, even if the psychological space is Euclidean, there are good reasons in favor of a non-standard multiplicatively-weighted determination of distances, under which conceptual regions could be non-convex.

Finally, even if none of the above problematic possibilities were the case, Gärdenfors' convexity constraint is brought into question by his own characterization of the working of conceptual spaces. The problems could be summed up as follows. [I] Some of the allegedly convex properties of concepts are not convex, as happens with those associated with the *shape* domain, and it is not clear how they could be represented in a convex way. [II] The conceptual region resulting from the combination of two –or more– convex properties belonging to the same domain may be non-convex, and the same happens for its associated concept. [III] The space resulting from the covariation of different convex regions could be non-convex, as a theoretical possibility that needs for more empirical scrutiny. [IV] Gärdenfors' definition of verb meaning, as vectors from one point to another, is not compatible with a definition of properties in terms of convex regions.

Based on all this, I conclude that since the convexity requirement can be detached from the theory with little impact on its explanatory capacities, the mandatory character of the convexity for regions in Gärdenfors' conceptual spaces should be rethought. Additionally, once the convexity requirement is given up on, it is possible to speak of a conceptual space theory without mentioning the shape of the underlying conceptual regions; and, in consequence, to conceive conceptual spaces merely in terms of a contextualist geometric articulation of the prototype theory of concepts.

4.4. Summary

This chapter has been devoted to describe the main notions and theses of the conceptual space theories (as a general framework for the similarity-based geometrical representation of concepts), paying special attention to the case of Gärdenfors' conceptual spaces. With this aim in mind, I have first reviewed how the notion of similarity may be characterized in this type of approach –namely, as a measure inversely proportional to distance–, and how Voronoi diagrams can be used both to determine the shape and boundaries of conceptual regions, and to carry out categorization tasks.

The second part of this chapter has been focused on Gärdenfors' theory of conceptual spaces, which is possibly the most influential perspective at present. There I have provided a detailed introduction of Gärdenfors' conceptual spaces, examined his motivations, major assumptions, theses and applications, and explained the role played by the main elements of his theory (i.e., *dimensions*, *properties*, *concepts* and *domains*). One of those basic theses in Gärdenfors' view was the *convex* structure of the regions associated with properties and concepts. Next, in the third part of the chapter, I have tried to prove that the convexity constraint is unnecessary –from a theoretical perspective–, and problematic –with regard to the working of Gärdenfors' theory–.

On this basis, my view is that the convexity constraint may be detached from the conceptual space theory with no impact on its cognitive efficiency or on its explanatory power, because the account of most phenomena (e.g., categorization, concept acquisition and change, communication, conceptual vagueness, etc.) relies on THESIS V, and not on the convex –or non-convex– shape of conceptual regions. Additionally, if the convexity constraint is abandoned, then Gärdenfors' conceptual space theory is reduced to a contextualist geometric characterization of the prototype theory, so this will be my position in chapters 5 and 6 of this thesis, where I speak of conceptual spaces with no mention to convexity (nor to any other compulsory constraint on the geometry of conceptual regions).

Finally, I devote chapter 5 to the investigation of issues seldom addressed in the conceptual spaces literature. Firstly, I hold that two different notions of concept must be distinguished (i.e., *stored concept* and *instantiated concept*), which has consequences both for the presumed cognitive architecture, and for the ontology of concepts. Then I also claim that, under the assumption that concepts are always context-dependent, concepts lack persistence and have not representational character. Lastly, in chapter 6 I deal with the circularity problem that emerges in any empiricist theory of concepts, when trying to account for how the most basic constituents of concepts are acquired.

This page intentionally left blank

Chapter 5: Concepts and categorization processes in a contextualist framework: An anti-representational view

> *The function by which we thus identify a numerically distinct and permanent subject of disclosure is called* CONCEPTION*; and the thoughts which are its vehicles are called* concepts. –
> William James (1890, vol. I, p. 461)

5.1. Introduction

People have beliefs, and those beliefs are not stable and persistent, but provisional and variable. It is not strange that my current beliefs about Austria are different from what they were five years ago. At that earlier time those beliefs could consist in that Austria is a German-speaking and medium-sized country in Central Europe; while now, after living for some time in Salzburg, they could be –in addition to those I had five years ago– that Austria is also a green and beautiful land, with a rainy weather. (Or, in other words, my concept of AUSTRIA today may be different from my concept of AUSTRIA five years ago.) Obviously, beliefs can change from time to time by virtue of the new environmental and linguistic information available to the subject, thus any theory of concepts should be able to provide an account for it.

However, revision of beliefs –or conceptual change– is not the only way by which the same concept may vary. In particular, other very common way through which concepts are thought to change is contextual dependence. Remember that, as said in section 1.4, two main approaches may be distinguished regarding the degree of context-dependence of concepts. On the one hand, the traditional view –sometimes also called *invariantism* (Machery 2009)– identifies concepts with cores of knowledge stable across individuals and time, which allows to explain both the accumulation of knowledge about categories, and our ability to communicate with other subjects. On the other hand, *contextualism* is the second main view, according to which many concepts are context-dependent construals created on the fly for each particular occasion (Barsalou 1993; Sperber and Wilson 1995; Carston 2002; Prinz 2002; Malt 2010), which would explain our adaptive behavior to changing environments. For my part, as I have already said in chapter 1, I ascribe to the contextualist thesis and –in particular– to Casasanto and Lupyan's *ad hoc* cognition framework, which is located in the scope of radical contextualism.

Looking back to the first issue, the similarity-based space theories of concepts have been sometimes blamed for not explaining the phenomena of belief revision (Gauker 2007). Those criticisms, although not always well founded, usually arise from an inadequate distinction between two distinct senses of concepts –associated with two different facets in their life cycle (i.e., storage and instantiation)–, and from a poor characterization of the inner workings of the adopted approaches. Due to this, in the present chapter I hold that a conceptual space theory can be conceived in a way such that concepts are not static representations within a conceptual hyperspace, but the result of instantiation processes that are specific of each particular occasion.

With that aim in mind, and after showing that the contextualist *ad hoc* cognition framework can be characterized in terms of a similarity-based space theory of concepts articulated by means of prototypes, I argue for the need–in a framework like this– (i) to shift the focus from conceptual regions to prototypes, and also (ii) to distinguish between two different senses of concept (Hernández-Conde 2017a). Indeed, once concepts stop being identified with the regions, and the notion of concept as stored information is differentiated from the one responsible of its external manifestation (in categorizations, inferences, etc.), the doubts regarding the possibility of conceptual change do not arise anymore. My exposition will prove that this is so by describing in detail how –in a proposal like mine– concepts are acquired, how conceptual change / revision may happen, and how previous versions of a concept could be stored as a historical series registered and organized under a same mental file.

After setting those goals, in section 5.2 I present Casasanto and Lupyan's *ad hoc* cognition framework –as a concrete exemplification of the contextualist view–, according to which there would be no context-independent concepts (i.e., all concepts are *ad hoc* concepts construed on the fly when they are instantiated, and only exist when they are applied in categorizations, communication, inferences, etc.) There I argue for the need to distinguish between the prototype of a concept and its associated conceptual region in the psychological space, and I will point out that concepts should not be identified with the conceptual regions, but with the prototypes (and the latter only in a very particular sense). Then, I show that the *ad hoc* cognition framework could be characterized through a prototype theory articulated by means of a similarity-based conceptual space, and I identify four possible sources of contextual dependence, namely, relevant concepts, kind of metric, importance of dimensions, and weighting of the considered concepts.

Next, in section 5.3, I explain why a proposal like mine (i.e., a contextualist approach characterized by means of prototypes and similarity-based geometric spaces) leads to the necessity of distinguishing two distinct notions of concept, namely *concepts as storage* and *concepts as instantiation*, which could be associated with different facets in the life cycle of a concept. On the one hand, the *stored concept* registers the location of the prototype associated to that category. In fact, that location is all the information needed to be persistently kept by the mind about a concept for its later instantiation. The registering of successive versions of that information (i.e., prototype locations) under the same mental file guarantees the continuity of the concept, and its ability to accumulate new information about that category. On the other hand, the *instantiated concept* is the result of those cognitive processes where the concept is applied (i.e., categorizations, inferences, etc.) Therefore, the instantiated concepts are the responsible of the external manifestation of a concept, and may be identified with the notion of concept proposed by (radical) contextualism and, consequently, with Casasanto and Lupyan's *ad hoc* concepts.

My aim in section 5.4 is to discuss the relation of mutual dependence that exists between the notions of stored concept and instantiated concept. First of all, I will make explicit the structure of the life cycle of a concept assumed in my work, whose three main stages are associated to the acquisition, revision and application of concepts. On this basis I hold that the life cycle of a concept is not linear but circular, and that storage and instantiation play two very different roles in it. In spite of this, my point is that there is a strong mutual dependence between both notions. On the one hand, the instantiated concept depends on the stored concept because part of the information accumulated in the stored concept about a category is used by the cognitive processes which instantiate such a concept. On the other hand, because previous categorizations of objects as *C* –resulting from past instantiations of the concept *C*– can be used by subsequent revision processes that change the stored concept associated to *C*, it may be said that stored concepts depend on instantiated concepts.

Then, in a discussion part (i.e., section 5.5), I compare my position with other similar views, like that of Machery, and –primarily– with Barsalou's and Prinz's, and try to elucidate the main commonalities and differences between my perspective and those of these authors. In section 5.5 I also claim that one of the main advantages of my approach is that it gathers together virtues both from contextualism and invariantism. On the subject of contextualism, my proposal suitably articulates a framework –that of the *ad hoc* cognition– compatible with the evidence against the existence of definitions –or conceptual cores–, in which *instantiated concepts* are construals produced on the fly from a set of context-specific cues. By virtue of this, the approach advocated in this doctoral thesis can account for our adaptive ability to changing environments. With regard to the subject of invariantism, my view is able to explain how, although instantiated concepts are context-dependent, *stored concepts* are stable enough to be the means by which new persistent knowledge may be accumulated about categories. Lastly, since I do not assume anything susceptible to be identified with a stable set of membership conditions, my approach works perfectly well without resorting to the idea of definition –or conceptual core–.

In section 5.6 I argue that, if concepts are assumed to be context-dependent –as hold by contextualism–, then instantiated concepts are non-persistent mental events and, as a consequence, are not a representation of their associated categories. Firstly, I show that instantiated concepts are the result of cognitive processes, not in the sense of *products* (i.e., persistent psychological entities stored in mental structures), but in the sense of *phenomena* (i.e., as that which happens when something is classified under a certain category). Thus, my point will be that instantiated concepts are non-persistent mental events that occur in the mind. Next, I maintain that, if representations are standardly described as relatively stable objects that codify information, then it may be concluded that instantiated concepts cannot be representations, since they do not fulfill the *minimal persistence* condition that is always demanded of the notion of representation.

Finally, in section 5.7 I will hold that the necessity of distinguishing between stored concepts and instantiated concepts may be generalized to any position that assumes a contextualist view of concepts. The idea is that, because none of the key elements in my argument decisively depends on the chosen theory on the structure of concepts, nothing prevents the same considerations to be applied in other approaches, as long as a contextualist view of concepts is assumed.

5.2. A conceptual space approach to radical contextualism

In this section I introduce Casasanto and Lupyan's radical contextualist approach, according to which concepts always depend on context. On their view, concepts are produced on the fly –from a set of contextual cues[1]– whenever they are used for categorizing, communicating, drawing inferences, etc., and only exist when they are instantiated for these sorts of tasks. Withal, the main weakness of Casasanto and Lupyan's view is that it lacks a proposal for articulating it within a theory on the structure of concepts. My aim here is to show that the *ad hoc* cognition framework can be characterized by means of a similarity-based approach to the prototype theory of concepts.

5.2.1 The ad hoc cognition framework

Casasanto and Lupyan (2015) have proposed an appealing and daring thesis: *there are no context-independent concepts* –that is, *all concepts are ad hoc concepts*–. They also argue that the seeming stability of concepts is merely due to commonalities across their different instantiations but that, in fact, there is nothing invariant in them[2]. On their view, which is based on Wittgenstein's discussion of *family resemblances* for the term "game" (according to which there is nothing in common to all the activities we call "games" (Wittgenstein 1953, §66-100)), the phenomenon identified by Wittgenstein is completely general, that is, it extends to every possible concept. And since the *core*[3] of a concept is conceived as those properties common to every object categorized under that concept (i.e., properties which are essential to that category, independently of the considered context), then it cannot be drawn a boundary between a concept's core and its "periphery", by virtue of the impossibility of identifying the core of no concept[4]. Casasanto

[1] Although different kinds of context-dependent information (CDI) may be distinguished, as Barsalou does when he separates CDI activated by the current context and CDI activated in recent context (Barsalou 1989), and even though I commonly agree with that sort of distinctions, none of them will be considered in my work, where I will only use the general notion of *context-dependent information*.

[2] Casasanto and Lupyan's position is clearly inspired by Barsalou's (1987; 1989) works.

[3] According to the distinction between a concept's *core* and its *identification procedure* (Osherson and Smith 1981; Armstrong *et al.* 1983).

[4] As said in section 2.2.1, Wittgenstein's work is usually interpreted as a critique of the classical theory of concepts –in particular, of the view that concepts are definitions–, since there are no necessary and sufficient conditions that determine the classification of something under a given category. And, although some answers to this issue suggest that the replacement of the classical theory by the prototype theory overcomes this issue (Rosch and Mervis 1975), on my view the difficulties identified by Wittgenstein are not restricted to the classical theory, but affect any invariantist approach to the notion of concept –including the invariantist interpretations of the prototype theory–.

The reason is that the invariantist view considers that that which determines the categorization of an object under a particular concept must be something stable in time and shared between individuals. Indeed, for the case of an invariantist prototype theory, that stable information would be the location and Lupyan convincingly argue that it is necessary to abandon the idea that concepts have stable –or default– cores accessed by people when they instantiate those concepts. Or, in other words, that there is nothing invariant in concepts, so there is no set of stable and context-independent properties accessed whenever subjects instantiate a concept.

In agreement with this, every instantiation of a concept would be produced on the fly from a set of contextual cues in an occasion specific manner. In particular, Casasanto and Lupyan (2015, pp. 553-557) distinguish three types of overlapping contextual information depending of the considered time scale: (I) *Brain activation dynamics*: the subject's cognitive state is always changing, as a result of its own brain activity, which entails a continuous reconfiguration of the cognitive system in function of its acts of perception and conception (i.e., in terms of the currently perceived inputs and instantiated concepts). (II) *Local context*: subjects instantiate concepts based on the cues received from their local contexts (i.e., physical, social, biological and neuro-cognitive), which has influence over the mental representations produced by those subjects. (III) *Experiential relativity*: persons are exposed to different linguistic, cultural or bodily experiences, and that may explain their distinct conceptualizations of time, space, movement, color, morality, etc.

Based on this, Casasanto and Lupyan maintain that, given that the subject's cognitive state is a part of the context, and considering that the brain is continuously changing, this implies that concepts are inherently variable. Hence, if Casasanto and Lupyan are right, concepts would only exist when they are instantiated[5] (i.e., when they are applied by a subject in categorizations, communication, inferences, etc.), and it is for that reason that they sum up their view as follows:

> Concepts are not something we *have in* the mind, they are something we *do with* the mind. (Casasanto and Lupyan 2015, p. 546)

For example, the instantiation of a concept when we categorize or make inferences is something we *do with* the mind, and not something we *have in* the mind. For my part, I sympathize with the view that where we "see" concepts, what there is in fact is the result of cognitive processes (i.e., categorization, comprehension, inference, etc.) However, after asserting that *concepts are not something we have in the mind, but something we do with the mind*, Casasanto and Lupyan focus their work on the instantiation of concepts, leaving aside the issue of which cognitive structures might ground those instantiations. Indeed, what they say regarding the information required to instantiate a concept is too vague to be an explanation of how that process happens:

of the prototype associated with each category, so there would exist something in common to all the members of a given category, namely the fact that all of them fall within the region associated to that concept. But, that was precisely what Wittgenstein proved that did not happen for the case of the term "game", so Wittgenstein's remarks do not only call into question the classical theory, but also support contextualist versus invariantist approaches to concepts.

[5] From here on in my doctoral thesis I will adopt Casasanto and Lupyan's general position, so I will assume that concepts are cognitive tools used by the mind in categorization tasks. This is in line with the second view on the ontology of concepts referred in chapter 1, according to which concepts are mental abilities that allow individuals to discriminate members from non-members of a given category.

> We will use the term *concept* to mean a dynamic pattern of information that is made active in memory transiently, as needed, in response to internally generated or external cues. (…)
>
> Rather than a process of accessing a preformed package of knowledge, instantiating a concept is always a process of activating an *ad hoc* network of stored information in response to cues in context. (Casasanto y Lupyan *ib*.)

Therefore, in order to accept the theses of the *ad hoc* cognition framework, a characterization of the cognitive structures supporting the instantiation of concepts is demanded. The rest of this section is devoted to the issue of how Casasanto and Lupyan's approach might be articulated by a theory on the structure of concepts –and, more particularly, by the prototype theory–, paying special attention to the question of how the context-dependence of every instantiated concept may be put in place. Thus, and in line with the positions adopted till this point of my thesis, from here on I will try to show that the *ad hoc* cognition framework may be conceived in terms of a prototype theory of concepts developed by means of a geometric similarity space.

5.2.2 On the distinction between prototypes and conceptual regions

Advocates of conceptual space theories sometimes identify concepts with the regions that result from a Voronoi partition of the conceptual hyperspace –taking as starting point the prototypes of the concepts relevant in the considered context–. Indeed, that is Gärdenfors' point when he defines (natural) concepts[6] as sets of convex regions in a number of domains[7] (Gärdenfors 2000, p. 105)[8, 9]. By contrast, others have wondered what reasons could have the mind to draw boundaries between regions in a similarity space

[6] See section 4.2.2 in this doctoral thesis.

[7] This perspective is almost equivalent to that according to which concepts are represented by the boundaries around their respective categories (Ashby and Townsend 1986; Ashby and Gott 1988; Ashby 1992; Maddox and Ashby 1993), which sometimes has been referred as the *category boundary* approach to the structure of concepts (Goldstone and Kersten 2003).

[8] Churchland, in the case of his connectionist approach, also seems to attribute a strong ontological import to the existence of an underlying conceptual space –with regions, boundaries and so on– when he says that "there really is an abstract space and it really does come to contain prototypical points, similarity gradients, category boundaries or partitions, and a well-defined geometrical configuration that embraces all of them" (Churchland 1998, p. 16).

[9] Let me observe that, although Gärdenfors insists on the connection between his conceptual spaces and the prototype theory of concepts, prototypes are less essential for his theory than convex regions. First, he defines properties and concepts in terms of (convex) regions, not in terms of prototypes that could incidentally lead to the determination of their associated conceptual regions. Or, in other words, according to the conceptual space theory held by Gärdenfors, concepts are substantially defined as convex regions, which only contingently have an associated prototype. Second, only attributing a secondary character to prototypes it is possible an approach to concepts in terms of *convex* regions. This is so because if prototypes were predominant over regions –with regard to the nature of concepts– (as hold by me in section 5.2.2), then the only constraint required over the geometry of conceptual regions would be star-shapedness, not convexity. Hence, it is fair to say that in Gärdenfors' theory prototypes are only a means to an end, and that end are the convex regions that he identifies with properties and concepts.

(Gauker 2007). Although for reasons different from Gauker's, in this section I argue against the thesis that the mind draws boundaries delimiting concepts. Here my point is that there are no reason why having a concept involves drawing its boundary. More specifically, I will hold that –according to the conceptual space theory– it is possible to possess a concept and not having to determine its associated boundary and/or conceptual region[10].

In this regard the key question is: What must be identified with the notion of *concept* in a similarity space approach based on the prototype theory? There are two possible answers to this question: (i) the concepts are the *prototypes*; and (ii) the concepts are the *conceptual regions* produced from those prototypes. In principle, the issue seems innocuous, to the point that in many cases concepts are identified either with the prototypes or with their associated regions indistinctly. The reason for this is that prototypes and conceptual regions are interdefinable since: (a) given a certain set of prototypes, the regions associated with their respective concepts can be determined; and (b) given a set of conceptual regions –resulting from a Voronoi tessellation of the conceptual space–, the location of their respective prototypes may be established.

Here my point is that regions and prototypes are distinct things that play very different roles. On the one hand, *conceptual regions* play mainly an explicative function, since it is easier to say briefly that «an object *o* falls within the region associated to a concept *C*», than to say that «the distance between the object *o* and the prototype associated to the concept *C* is smaller than any of the distances between *o* and the prototype of any other concept distinct from *C* (and relevant in the considered context)». On the other hand, *prototypes* are the notion which intervenes in cognitive processes, such as concept formation and categorization, and there are significant reasons which support this statement:

— That which results from the generalization of a set of (tentative) examples of a given category is a prototype, not a region. Indeed, conceptual regions only arise from the evaluation of the distances between all the points of the conceptual hyperspace, and the prototypes of the relevant concepts.

— The application of conceptual regions in categorization tasks is both unnecessary and inefficient: (1) It is *unnecessary* because in order to categorize an object only the location of the relevant prototypes is needed[11]. (2) It is *inefficient* –both in terms of memory and/or processing– because it compels, either to store the concept associated to every point of the conceptual hyperspace, or to store the whole boundaries and calculate the region within which the considered object is located.

[10] Since in this kind of discussion boundaries and regions are interdefinable and interchangeable notions (inasmuch as, given a certain region its boundary is defined; and given a particular boundary, its delimited region is fixed), throughout the rest of this section I restrict my discourse to the case of conceptual regions.

[11] Obviously, those boundaries might be determined from the very same input used to categorize an object under a given concept (i.e., from the locations of the prototypes of the relevant concepts). Nonetheless, all those calculations are not needed and, consequently, very possibly not done.

Hence, it is an error to attribute to the conceptual regions a persistent and strong ontological sense –as actually existing entities present in our minds– seeing that their function is merely explicative, and that they do not intervene in any significant cognitive task.

Therefore, in my view concepts are associated with prototypes, and not with conceptual regions produced from those prototypes[12]. In fact, my point is that the only information required to be stored by the mind about concepts is the location of their prototypes –and not their associated regions and/or boundaries–. This is crucial in order to hold that a prototype theory of concepts may articulate the *ad hoc* cognition framework, since if the information stored in memory were the regions, then it could not be said that the (instantiated) concepts depend on context. Indeed, under a contextualist view of concepts a context-independent determination and storage of the region and/or boundaries associated to a concrete concept is *useless*, because that region and boundaries will depend on information that varies from context to context –as shown in section 5.2.3–. In consequence, if concepts were the regions and/or boundaries, then every time the context changes[13] a whole recalculation of the boundaries would be needed, so it would be nonsense –both from a psychological and a computational point of view– to store information that will be rarely reused.

All in all, my suggestion is that it is necessary to shift the focus from conceptual regions to prototypes –or, in other words, from boundaries to prototypes–.

5.2.3 *A similarity-based model for ad hoc cognition*

Let us see now how a contextualist framework –like that of the *ad hoc* cognition– may be articulated by means of a similarity space theory of concepts. As said in section 4.1.3, conceptual space theories define similarity as a measure inversely proportional to distance. In particular, the distance between two objects (and/or prototypes of concepts) *a* and *b* within an *n*-dimensional space was the following:

$$d(a,b)=\left(\sum_{i=1}^{n} w_i \left| f_i^{[a]} - f_i^{[b]} \right|^p \right)^{1/p}$$

where $f_i^{[o]}$ is the value of the *i*-th dimension of the object *o*; w_i is the weight assigned to the *i*-th dimension; and *p* determines the kind of metric.

[12] Anyhow, it could be argued that there is a sense in which concepts may be identified with regions and boundaries, even though the unique information stored by the mind about categories was the location of prototypes. The idea underlying that critique could be that the instantiation of a concept leads *implicitly* to the application/determination of its associated region since: (a) the instantiation of a concept produces the same result that would be produced from applying the information about the whole region in that categorization task; and (b) if instantiations were carried out for all the points of the geometric space, they would result in the regions and boundaries of all the considered concepts.
Here my answer is that none of the previous considerations should be confused with an *effective* application/determination of the regions –or boundaries– associated to those concepts. On the contrary, they are mere useful ways of providing a high-level description of what is happening in such instantiations, and not a true explanation of how those cognitive processes really work.

[13] And, according to many contextualists, context is continuously changing.

Additionally, as said in the previous chapter, distances may be differently weighted, what led me to introduce the notion of distance-of-comparison $d_C(o,P_C)$ between an object o and a concept C –with prototype P_C–, which was expressed as follows:

$$d_C(o,P_C) = u_C\, d(o,P_C)$$

where u_C is the weighting of the distances from the prototype of C (i.e., P_C).

According to this view, an object o will be categorized under a concept G if the distance-of-comparison between o and G (i.e., $d_G(o,P_G)$) is less than the distance-of-comparison between o and any other relevant concept in that context. Or, in other words, if C is the set of relevant concepts in the considered context, then when $\forall C \in \mathsf{C}$ it is true that $d_G(o,P_G) < d_C(o,P_C)$, the object o will be categorized under the concept G. It is in this kind of cognitive process where the *instantiation* of a concept occurs, which consists in the evaluation of the similarities of a particular object –or concept– with regard to the set C of relevant concepts in that context[14, 15].

Thence, inasmuch as distances –and similarities– are a function of the parameters p, w_i and u_C, and given that the categorization of an object also depends on which the relevant concepts are (i.e., set C), it can be said that there exist at least four contextual factors that can affect the instantiation of every concept in a characterization of the *ad hoc* cognition framework like this:

— the instantiated concepts –set C of relevant concepts–,
— the kind of metric –parameter p–,
— the importance of dimensions –weights w_i–, and
— the significance of concepts –weights u_C–.

Relevant concepts

First, a categorization process will produce different partitions of the conceptual space depending on the set C of concepts relevant in the considered context (i.e., depending on the locations of the pertinent concepts), which will lead, consequently, to distinct instantiations of those concepts.

Consider the following example. Let be a subject S whose default conceptual space for the case of a categorization process of citruses is that shown in Fig. 5.1a, where the hori-

[14] If such a process happened for all the points of the conceptual hyperspace, that would produce a whole partition of the conceptual space in regions susceptible of being identified with the instantiations of all the concepts in C.

[15] Sometimes it is argued that most approaches are not really full-blown categorization models, since they merely focus on the last two steps (i.e., similarity calculus and decision making) of a three-stage process, and say nothing about the first stage (i.e., selection of relevant concepts) (Barsalou 1990; Machery 2009). On my view, that is a valid critique, since a fully developed model has to explain which concepts are considered in any particular categorization context. Indeed, my proposal would fall within that kind of not-full-blown approaches, because I have not delved into the details of how concepts are selected. Even though I am fully conscious of the significance of that first step, that issue lies within the context sphere and, as I have said previously, the problem of context (i.e., its characterization, maintenance and application) is outside the scope of this doctoral thesis.

zontal axis might be identified with the *color* dimension and the vertical axis may be identified with a mixture of *texture* and *shape*.

However, if the subject *S* were in a context where he did not expect that the fruit that grows in a tree could be a lime (perhaps because *S* is in a place where he knows that there are not lime trees, or because its presence there is quite rare), then LIME might not be a relevant concept in that categorization process. In that case, any object previously categorized as a lime would be now classified under the concept LEMON (see Fig. 5.1b). A similar phenomenon may happen if GRAPEFRUIT did not belong to the set C of relevant concepts, and instead TANGERINE was an element of C (see Fig. 5.1c); or if the subject *S* thought he was facing fake lemons (e.g., plastic, wooden or painted lemons) (see Fig. 5.1d).

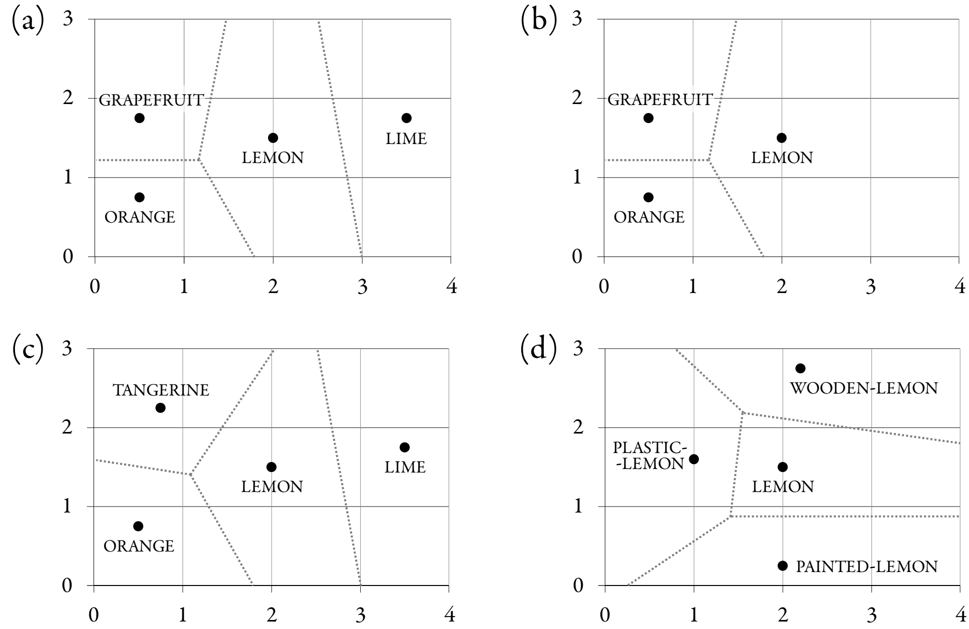

Fig. 5.1. Example of contextual dependence of concepts due to the set of relevant concepts, for a categorization process of citruses where abscissas may be identified with *color*, and ordinates with a mixture of *texture* and *shape*. (a) Default context with prototypes of the concepts LEMON, ORANGE, GRAPEFRUIT and LIME located in the coordinates (2,1.5), (0.5,0.75), (0.5,1.75) and (3.5,1.75), respectively. (b) Context where the concept LIME is not relevant. (c) Context where the third relevant concept were not GRAPEFRUIT, but the concept TANGERINE, located in (0.75,2.25). (d) Context where the relevant concepts were LEMON, PAINTED-LEMON, PLASTIC-LEMON and WOODEN-LEMON, the last three located in (2,0.25), (1,1.6) and (2.2,2.75), respectively.

KIND OF METRIC

Nonetheless, it might happen that there exist two distinct contexts $\mathcal{H}$ and $\mathcal{I}$ such that their sets of relevant concepts were the same (i.e., such that $C_{\mathcal{H}}$=$C_{\mathcal{I}}$, where $C_{\mathcal{X}}$ represents the set of concepts relevant in the context $\mathcal{X}$), but whose metrics were not identical. In that case, different metrics will produce, even for the same set of prototypes, distinct partitions of the conceptual space and, consequently, different instantiations of those concepts.

Now, consider again the instantiation represented by Fig. 5.1a whose metric was Euclidean –instantiation reproduced in Fig. 5.2a–. This could be the case of a context where the subject is so used to classifying citruses according to the dimensions of *color* and *texture-shape* that his perceptual and cognitive system jointly processed both dimensions.

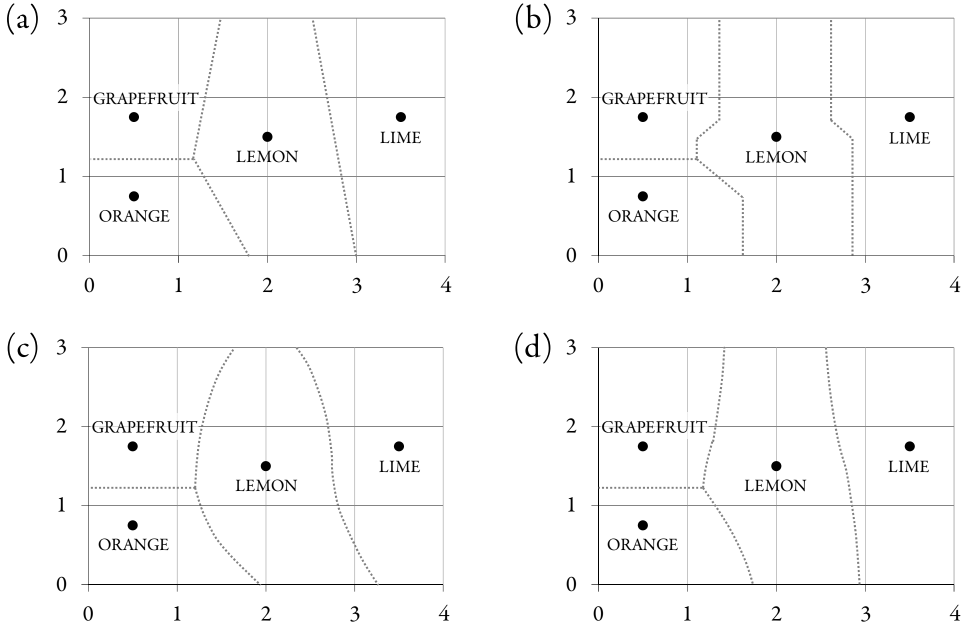

Fig. 5.2. Example of contextual dependence of concepts due to the kind of metric, for a categorization process of citruses with the same set of relevant concepts as the one in Fig. 5.1a, located in the coordinates (2,1.5), (0.5,0.75), (0.5,1.75) and (3.5,1.75), respectively. (a) Context with Euclidean metric (*p*=2). (b) Context with city-block/Manhattan metric (*p*=1). (c) Context with higher-order metric (*p*=3). (d) Context with optimal metric for integral dimensions (Handel and Imai 1972) (*p*=1.7).

By contrast, if the context were such that the dimensions of *color* and *texture-shape* were separately processed[16] (perhaps because the subject is encouraged to attend to the individual differences in those two dimensions; or at some previous time when that subject was not used to doing that task), then the applied metric might be the city-block (i.e., parameter *p*=1) (see Fig. 5.2b). Indeed, there is empirical evidence that the selective attention to the considered dimensions may change how similarity relations are deter-

[16] Remember that, as said in section 4.2.2, dimensions are integral if they are processed in an unanalyzable way, and assigning a value to one of them requires giving a value to the others (Garner 1974). When dimensions are not integral, then they are separable.

mined, so dimensions commonly integral (with parameter p=2) can be evaluated separately (parameter p=1), and vice versa (Melara *et al.* 1992)[17].

Obviously, instantiations would be different for other kinds of metric (see Fig. 5.2c and Fig. 5.2d for metrics with parameter p=3 and p=1.7, respectively).

Importance of dimensions

However, even though two different contexts shared the same set of relevant concepts and also the same kind of metric, if the importance received by the dimensions constitutive of the underlying conceptual hyperspace were distinct –in the limit, some weights could be equal to zero–, that would produce different instantiations of those concepts.

Look again at the instantiation shown in Fig. 5.1a, whose metric was Euclidean and where all dimensions are equally weighted (that instantiation is reproduced in Fig. 5.3a).

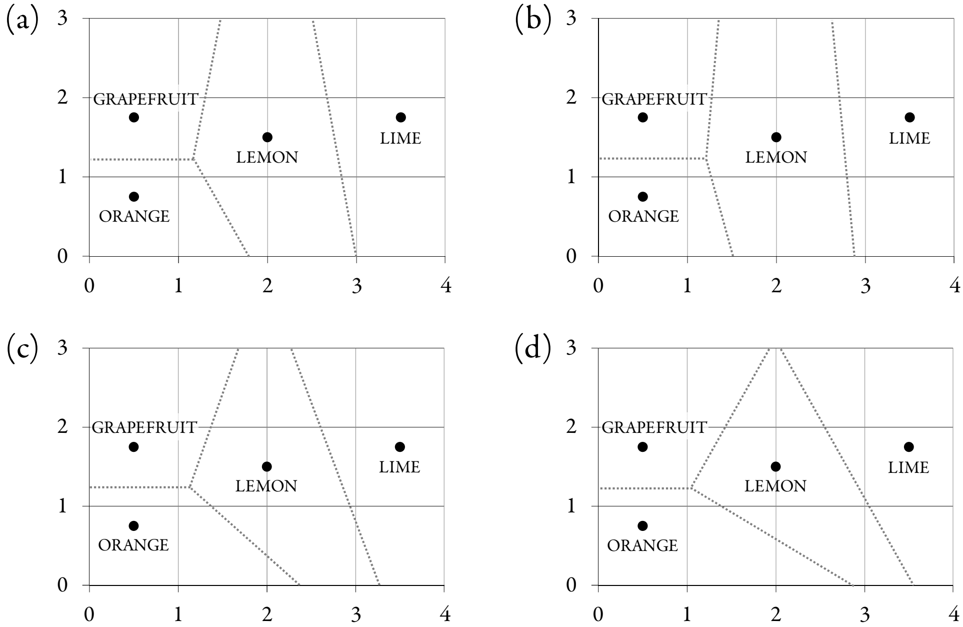

Fig. 5.3. Example of contextual dependence of concepts due to the importance of dimensions, for a categorization process of citruses with Euclidean metric, based on the *color* –horizontal axis– and a mixture of *texture* and *shape* –vertical axis–. (a) Default context with equally weighted dimensions [weights (1,1)]. (b) Context where *color* had twice the importance of the mixture of *texture* and *shape* [weights (2,1)]. (c) Context where *texture* and *shape* had twice the weight of *color* [weights (1,2)]. (d) Context where *texture* and *shape* had thrice the importance of *color* [weights (1,3)].

[17] In the same line Nosofsky (1987) proved that, if the difference of objects along their constitutive dimensions is very small, then a parameter of p=2 provides a better fit, even though such dimensions were separable. For a review of other cases where the same sets of dimensions can be distinctly processed –e.g., as a whole or separately– by the very same subject, see Goldstone and Steyvers (2001).

Consider now the case of a context where the subject watches the scene from a certain distance, by virtue of which his perception of the texture and shape of objects is not too accurate. In that case, the subject's cognitive system might overweight the dimension of *color*, for instance, assigning to it twice the weight of the mixture of *texture* and *shape*, which would produce a different instantiation of the considered concepts (see Fig. 5.3b). Alternatively, it could happen that –in another context– the *color* dimension had little importance (for instance, if the subject is in a dark environment, where the hue of color cannot be clearly distinguished; or if he was in a context of unripe fruits, where all of them were –more or less– greenish). In such a case, the mixture of *texture* and *shape* might have twice or thrice the importance of the *color* dimension, resulting in other two different instantiations of those concepts (see Fig. 5.3c and Fig. 5.3d, respectively).

Significance of concepts

Lastly, it could happen that although all the previous factors were the same in two particular contexts, the significance of concepts were not equal in both situations. In such a case, the distances-of-comparison (that are used in categorizations) would be differently weighted in each context, and that would produce distinct instantiations of the relevant concepts.

Let's consider again the conceptual space represented by Fig. 5.1a –where the four instantiated concepts (i.e., LEMON, ORANGE, GRAPEFRUIT and LIME) were equally weighted–, which is reproduced in Fig. 5.4a.

But, context could be such that concepts were distinctly weighted according to: (i) the relative frequencies of the examples observed in the subject's life course; and/or (ii) the subject's interests and/or expectations in the considered context. For instance, in the case of a weighting based on frequencies, if weights were (1.1,1.2,1,1)[18] (that is, if orange is the most frequent citrus, and lemon is the second one, equally followed by grapefruit and lime), the instantiated concepts would be those shown in Fig. 5.4b. By contrast, if the subject *S* works in a production line of lime nets where most of the citruses are limes, even though sometimes unripe lemons also appear, the subject *S* might be especially sensitive to limes, and slightly less sensitive to lemons –so that the weights of concepts were (1.3,1,1,1.5)– (see Fig. 5.4c). Finally, a fourth possible context might be one in which oranges and lemons were equally –and significantly– overweighed regarding grapefruits and limes, which would happen for the quadruple of weights (2.5,2.5,1,1) (see Fig. 5.4d)[19]. That would be the case if the subject –a child, for example– had been exposed to a very small number of grapefruits and limes; or, in other words, if the majority of citruses seen by the subject had been oranges and lemons.

[18] These weights –and all the other weights that will appear in this subsection– are relative to similarities, that is, they are the inverse of the weights u_C (associated to distances) that appeared in the multiplicatively weighted scheme shown at the beginning of section 5.2.3.

[19] Observe that the result shown in Fig. 5.4d of these low weightings for the concepts GRAPEFRUIT and LIME, as well as that represented for the concept FOX in Fig. 4.7b, are functionally equivalent to the result of the rule-plus-exception hybrid model proposed by Nosofsky and collaborators (1994).

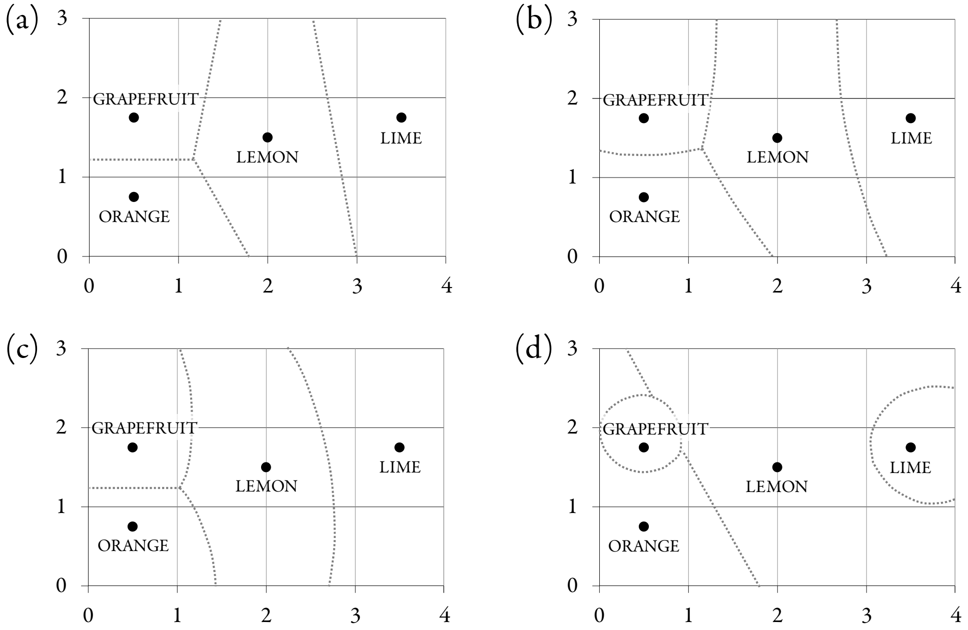

Fig. 5.4. Example of contextual dependence of concepts due to the significance of the relevant concepts, for a categorization process of citruses with the same set of relevant concepts as that in Fig. 5.1a. (a) Default context with equally weighted concepts [weights (1,1,1,1)] (associated to LEMON, ORANGE, GRAPEFRUIT and LIME, respectively). (b) Context with concepts weighted by their relative frequency [weights (1.1,1.2,1,1)]. (c) Context for a worker in a production line of lime nets [weights (1.3,1,1,1.5)]. (d) Context for a child who had been exposed to a small number of examples of grapefruits and limes [weights (2.5,2.5,1,1)].

* * *

Obviously, all these four context-dependent factors will not usually intervene individually in the instantiation of concepts –as discussed up to this point–, but all of them as a whole –in the way represented by Fig. 5.5b–.

All in all, each new instantiation of a concept in a particular context can be different, since the relevant concepts, the kind of metric and the importance of dimensions and concepts may vary from context to context.

Lastly, a prototype theory of concepts (conceived in terms of a geometric similarity space) can provide a successful account of Casasanto and Lupyan's main thesis, namely, that all concepts are *ad hoc* concepts –or, in other words, that the instantiation of every concept depends on the context where such an instantiation happens–.

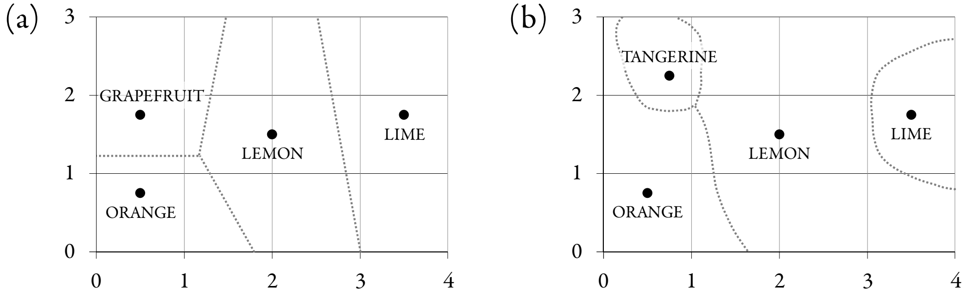

Fig. 5.5. Example of combined contribution of four contextual factors (i.e., *relevant concepts*, *kind of metric*, *importance of dimension*s, and *significance of concepts*), for a categorization process of citruses. (a) Default context with prototypes of the concepts LEMON, ORANGE, GRAPEFRUIT and LIME located in the coordinates (2,1.5), (0.5,0.75), (0.5,1.75) and (3.5,1.75) –respectively–, for a Euclidean metric where all the concepts and dimensions were equally weighted. (b) Alternative context where the concept GRAPEFRUIT is not relevant, but the concept TANGERINE –located in (0.75,2.25)–; for a metric with parameter p=3, where the *color* dimension had twice the importance of the mixture of *texture* and *shape* [weights (2,1)], and concepts had different significance [weights (2.5,2.5,1,1)].

5.3. *Two-faceted concepts*

Hitherto, two distinct notions of concept have been tacitly used in this chapter. The first is that associated with the information persistently stored by our mind about a given category –I will call concepts in this sense *stored concepts*–; while the other is that referring to the result of those cognitive processes which apply part of the information stored about that category[20] in cognitive tasks (i.e., categorizations), for each particular context –I will call concepts in this other sense *instantiated concepts*–.

In this section I will hold that it is worthwhile to distinguish those different notions –which should not be confused–, and that they may be identified with two distinct facets in the life cycle of a concept.

5.3.1 What is "having a concept"?

However, before that discussion, and without trying to provide a definition of what a concept is, I will briefly sketch out how *having a concept* may be understood, and the main difficulties associated to the distinct approaches to this question.

A GENERAL ISSUE WITH THE IDEA OF "HAVING A CONCEPT"

According to Putnam (1970, 1975) and Kripke (1980), to be competent with a natural kind concept[21] cannot require knowing the conditions of application –or category mem-

[20] Together with information stored about other relevant categories in that context, and other contextual parameters (e.g., the kind of metrics) and weightings (e.g., of dimensions and concepts).

[21] Although Putnam and Kripke focus their discussion in natural kind terms/concepts, there is nothing in their arguments that prevents them from applying to the case of general concepts.

bership–, because in many cases we are competent without knowing those conditions –sometimes due to ignorance and others due to error[22]–.

However, as said by Margolis (1994), Putnam's and Kripke's line of argument may be used, not only against the classical theory, but also against the prototype theory. His point is that when the assumptions of these theories are made explicit, both of them share a same epistemological condition:

> (3) The classical theory
> (a) All instances of a category share a set of properties singly necessary and jointly sufficient for membership within the category.
> (b) Having a concept involves knowing the conditions of membership within the corresponding category.
>
> (4) Prototype theory
> (a) Category membership is a matter of having some sufficiently many properties that members of the category tend to have.
> (b) Having a concept involves knowing the conditions of membership within the corresponding category. (Margolis 1994, p. 78)

Under this view, both theories demand that subjects know the conditions for category membership –namely, condition (3b) / (4b)– so, if Putnam and Kripke are right, in the sense that the crucial fact is that people cannot –in general– specify the conditions for category membership, this would occur independently of whether such conditions are conceived in classical or in prototypic terms (i.e., in terms of a condition like (3a) or like (4a), respectively). In such a case, the prototype theory is no better off than the classical theory, since condition (3b) / (4b) is the same in both cases.

Even more, that very same conclusion will affect any other theory on the structure of concepts which shared a premise like (3b) / (4b). In fact, any invariantist theory on the structure of concepts will have to face Putnam's and Kripke's line of argument, because what underlies the condition (3b) / (4b) is the assumption that there exists a set of *invariant* –and, consequently, free from the problems of ignorance and error– conditions for category membership[23].

[22] For a description of the problems of ignorance and error, see section 2.2.1.

[23] The same line of argument by virtue of which the problems of ignorance and error become irrelevant, allows to easily explain the existence of failures of communication: because it is not possible to know the membership conditions of a given category, a subject can never be sure that her utterances have been properly understood by the other participants in a conversation. Nonetheless, it is precisely that point which makes communication into a problem *per se* for contextualism, since the contextualist has to provide an account of how that thing we call "successful communication" may happen without the help of a set of shared –and identical– concepts/meanings. (And what I have said about *successful communication* could also be said regarding other human activities such as *cooperation*, *agreements*, *language*, etc.) For my part, and even though I recognize that the latter is a crucial issue, its study lies beyond the aim of this doctoral thesis. Anyhow, the answer to that question might be based on the satisfaction of one's own expectations on the basis of (apparently) shared classifications of the same examples. (That could lead the subject to conclude that one of her own concepts *C* is close enough to the concept *C* of other subject –or set of subjects–.)

Despite this, the prototype theory may also be conceived in non-invariantist terms (in line with the radical contextualist view assumed in my work), and in such a case it would be beyond Margolis' generalization of Putnam's and Kripke's critique. The reason is that, according to radical contextualism, concepts require to be instantiated for each particular occasion –or context–, so it is meaningless to speak of conditions for category membership without an associated instantiation process. Or, in other words, condition (3b) / (4b) is irrelevant in a contextualist approach where those conditions of membership are dynamically produced for each particular occasion and, due to this, are constantly varying from context to context.

And, for the same reason it is also meaningless to speak of having a concept without considering a particular instantiation process. Under this view, and since the context may be continuously changing, the question of what is "having a concept" has no sense, and we should instead focus on the issue of competence with concepts in categorizations, inferences, communication, etc.

Contextualism, prototypes and the notion of "having a concept"

Anyhow, due to the difficulties described in the previous section and the tentative upshot towards shifting the focus to the affair of competence, we might wonder whether the notion of *having a concept* may be fixed from the field of competences. As said in chapter 1, we seldom face the same entity twice, and that is the reason why we depend on concepts in order to understand what is happening around us. Indeed, we are constantly classifying new objects under known categories, and attributing to them properties previously seen in other examples of those categories. Therefore, it might seem reasonable to think that we have a concept *G* if we are able to classify something under that category, or if we are able to carry out inferences and/or predictions about the objects categorized as *G*, or if we are able to communicate by using the word(s) associated to that concept[24].

However, as shown above, the determination of whether an object *o* belongs or not to a concept *G* is open to context in –at least– the four aforesaid factors[25]. This is the reason why a context-independent definition (or a context-independent identification procedure/algorithm) cannot be used in order to carry out the categorization of something under a particular class.

Thus, to the extent that the information stored of a concept is not enough –in the absence of a context– to classify something under that category, then the contextualist has a good reason to call into question that the mere information stored about a category may count as having that concept. Therefore, in the same way that Wittgenstein held that the

[24] Obviously, my perspective is also quite close to Barsalou's, when he argues in favor of reserving the term *concept* for the temporary constructions in working memory that control our behavior and performance in categorizations (Barsalou 1993, p. 34).

[25] Remember that the categorization of an object *o* under a concept *G* depends, not only on the information stored about *G* (i.e., the location of its associated prototype P_G), but also on the information stored about other concepts relevant in that context (i.e., the location of their prototypes –that is, the location of every prototype P_C such that *C* belongs to the set of relevant concepts C in that context–), together with other contextual factors (i.e., kind of metric, weight of dimensions and importance of the relevant concepts).

meaning of a word is –for a large number of cases, though not all– its use in the language (Wittgenstein 1953, §43), as part of a form of life, a contextualist might argue that the concept associated to a particular category is its application (e.g., in categorizations, inferences, communication, etc.) in a given context. Thence, it could be thought that for a contextualist the question of whether somebody has a concept is nonsense without an application –of that concept– in a certain context; as well as the question about the meaning of a word is nonsense for Wittgenstein, without a use of that word in a particular form of life. Consequently, under this view the contextualist thesis that «concepts are constructed *ad hoc* when they are instantiated» (i.e., are applied in a given context) could be better understood[26].

5.3.2 Concepts as storage

The first notion of concept is that associated with the information stored within our cognitive systems about a given category. From here on I will refer them as *stored concepts* –or *concepts as storage*–.

As stated above –see section 5.2.2–, in the case of a proposal like the one defended in my thesis (i.e., a prototype theory of concepts built over a geometric similarity space), the only information which has to be registered by the mind is the location of the prototype associated to each concept. Those locations were the only thing required to instantiate a concept within a particular context –that is to say, to determine the distances and similarities between a concept and any other object or concept–. Therefore, stored concepts are the information persistently registered by our minds about the location of the prototypes of their respective categories[27,28].

The notion of *stored concept* presents virtues commonly associated to the invariantist approach to concepts. The reason is that this first sense of concept may be thought to be persistently backed by a certain structure –either informational (e.g., record system, neural network, etc.) or physical (e.g., potential level, electrochemical gradient, etc.)–, which guarantees the continuity required in order to accumulate new information over time about a certain category. For instance, if as a result of subsequent executions of the learning processes new properties are identified as relevant features of a particular category,

[26] This is consistent with the minimal empirical commitment that the only thing we observe about concepts are those tasks they allow us to carry out (e.g., to categorize), and points to a view of concepts as mental abilities (i.e., as cognitive tools used by the mind in categorization tasks).

[27] Obviously, this is not too much information to be stored. In fact, and since not all the hyperspace structure has to be registered, a history of the past prototypes of each concept might be maintained in their associated mental files.

[28] When I hold that *stored concepts* are stable and persistent, I do not betray Casasanto and Lupyan's principle that "concepts, categories and meanings (...) only exist as theoretical abstractions" (Casasanto and Lupyan 2015, p. 543), since such a principle just applies to the notion of *instantiated concept* (as that which is responsible of the external manifestation of categories). Anyhow, Casasanto and Lupyan might not be disposed to accept my notion of *stored concept* –by virtue of their inclination towards connectionism and, consequently, their rejection of the notion of "mental file"–. However, that would be a divergence in how information is stored (e.g., in a mental file, in a neural network, etc.), but not a divergence with respect to the existence of stored information from which concepts are instantiated.

such new properties can be simply added under the same mental file[29] where the stored concept (i.e., location of the prototype) was registered by the mind. Hence, if the stored concept is conceived as a set of information (about a given category) registered under a given mental file, then the informational / physical structure that supports such a stored concept warrants the aforementioned continuity over time of the information maintained about categories. The advantage of this is that, from a radical contextualist approach, it is possible to explain a typically invariantist ability –to wit, the accumulation of new knowledge by individuals–.

Nevertheless, it would be an error to confuse the information stored about a category with the storage of the sense of concept responsible of its external manifestation –namely, in categorizations, inferences, etc.–, seeing that the notion of concept which intervenes in those phenomena is not the stored concept, but the concept instantiated in each particular context. In fact, although the stored concept is the starting point for any instantiation of a concept –which may take place in cognitive processes such as categorization–, the stored information of a concept cannot determine the output of those processes by itself, since additional contextual factors are involved. Remember that, as shown in section 5.2.3, the instantiation of a concept requires calculating the distances / similarities between the evaluated object and the prototypes of all the concepts relevant in that context; and that such a computation depends, not only on the kind of metric and the importance of dimensions, but also, and more importantly, on the locations of the prototypes of all those relevant concepts, and on their significance for the considered context.

By virtue of this, the information stored about a category should not be identified with the invariantist notion of concept –in the sense of body of knowledge stored in long term memory, and used by default in the processes that underlie our higher cognitive abilities (Machery 2009, p. 12)–. In fact, the main difference between that what I have called concepts as storage and Machery's default bodies of knowledge, is that the latter are assumed to be always relevant for the subject –regardless of the considered context–, what is hardly compatible with a radical contextualist approach like the one adopted in my work[30].

5.3.3 *Concepts as instantiation*

The second sense of concept is the result of cognitive processes such as categorizations and inferences, and it is the notion responsible of the external manifestation of categories. Hereafter I will refer them as *instantiated concepts* –or *concepts as instantiation*–.

[29] In the sense of Recanati's *mental files* (Recanati 2012). A quite similar notion is used by Margolis (1998) when he discusses the acquisition of concepts.

[30] It could be thought that Machery's view might be articulated in terms of a general –or neutral– context which worked by default. However, as argued by Casasanto and Lupyan, there are significant arguments against the thesis that such default context might exist. More specifically, Casasanto and Lupyan (2015, p. 554) mention three kinds of evidence: (a) The representations produced by people in response to linguistic stimuli are a function of their social and physical environment. (b) Thinking and language understanding are partially determined by the situation and perspective of people regarding space and time. (c) The efficiency of memory retrieval depends on the subject's state of motion.

THE INSTANTIATED CONCEPTS ARE *AD HOC* CONCEPTS

My first point is that the notion of *instantiated concept* may be identified with the *ad hoc* concepts proposed by Casasanto and Lupyan (2015), that is, they can play the role attributed to concepts by a radical contextualist approach. It should be stressed, however, that contrary to Casasanto and Lupyan's view –who do not specify how a proposal like theirs may be deployed by a theory on the structure of concepts–, the instantiated concepts proposed in this chapter are fully articulated in terms of prototypes and similarity spaces.

Besides, the sense of *ad hoc* concept set up by my instantiated concepts has significant differences with regard to the usual ways in which the former is conceived. Thus, and even though there is no unitary view of what an *ad hoc* concept is, most of the authors that employ this notion hold that the *ad hoc* concepts are representations pragmatically produced for each particular context –in tasks such as categorizations, inferences, comprehension, etc.–, and susceptible of being identified with mental entities present in the subject's working memory (Barsalou 1987; Carston 2002; Allott and Textor 2012). Against this position, my thesis is that *ad hoc* concepts (i.e., *instantiated concepts*, in my terminology) are just the result of psychological processes, characteristically produced in each specific occasion[31].

In an approach like the one here proposed, instantiated concepts can be thought to be *mental events* –not mental entities–, since they are that which happens at the end of the cognitive process that instantiate a concept (in a particular context). And, for instance, that which results –in the case of a categorization– is the classification of an object *o* under a category *C*. Obviously, that result could be registered under the mental file associated to the object *o*, which attributed the property *C* to the object *o* (i.e., which said that «*o* is a *C*» according to a past categorization of *o*[32]). At this point it could be argued that, because the property *C* is stored in a registry under *o*'s mental file, then *C* might be identified with a mental entity in the subject's mind. However, that registry should be viewed, not as the instantiation of the concept *C*, but as a mere static attribution of a label (i.e., that associated to the property of being a *C*). Indeed, a crucial difference is that while the instantiated concept depends on a context specifically produced on each occasion; the information stored about a past categorization is independent of context –or, at best, associated to a poorer context and, consequently, different from that which produced that instantiation–.

Now, since the idea of instantiated concept is much more slippery than the notion of stored concept, I think that it is useful –for the purpose of this section– to consider the

[31] Could *instantiated concepts* be identified with the cognitive processes which produce them? Here my answer is no. In my view that which unifies distinct objects classified –in different occasions– under the same category *C* is the fact that the information about *C* used in all those categorization processes was stored under the same mental file –even though the stored information might have changed from one categorization to another–.

[32] The storage of a past categorization of *o* (e.g., that of «*o* is a *C*») would allow the subject to quickly attribute the property of being a *C* to the object *o*, without having to carry out a whole new categorization process.

analogy that may be drawn between those two notions of concept and the case of genes. In particular, such analogy can be made on the basis of the distinction between genetic information and expressed gene, and of the following correspondence[33]:

— *Genetic information* –or *gene as* (unit of) *information*–: the genetic information may correspond with the notion of *stored concept*, because in the absence of a process which provides the context (i.e., the instantiation process in the case of concepts, and the expression process in the case of genes) both of them are mere data –or information– registered in a physical medium[34].

— *Expressed gene*: which corresponds with the notion of *instantiated concept*, since the expression of a gene (i.e., production of a protein and appearance of the phenotype associated to that gene) at a particular time may be modulated by specific factors that alter –or regulate– the expression process, which resembles the occasion-specific production of instantiated concepts on the basis of a set of contextual factors.

Here my point is that the instantiation –or expression– of a concept is as eventive as the expression of a gene. On the one hand, the expressed gene (in the form of the production / appearance of a certain protein / phenotype) is an event that results from an expression process. Therefore, the expression of a gene is something that happens, and not a kind of genetic entity produced from genes as units of information. On the other hand instantiated concepts are events that result from instantiation processes (in the form of categorizations, inferences, etc.). Likewise, the instantiation of a concept is something that happens, and not a kind of mental entity produced from information stored in memory about categories.

Thence, my thesis is that concepts as instantiation should be understood as events that result from their associated cognitive processes (e.g., categorizations, inferences, etc.), in spite of which they are the sense of concept responsible of the external manifestation of categories[35].

Finally, my claim is in line with Casasanto and Lupyan's lemma quoted in section 5.2.1 –according to which concepts are not something we have in the mind, but something we do with the mind–. In particular, I maintain that instantiated concepts are not something that exist (in the mind), but conversely, they are something that occur at the end of their respective instantiation processes[36].

[33] Thanks to Laura Nuño, whose comments at the 8th Research Workshop on Philosophy of Biology and Cognitive Science (Complutense University of Madrid, May 2018) led me to be aware of the parallel between the case of concepts and that of genes.

[34] Remember that, as said above in section 5.3.2, stored concepts may be thought to be backed by a certain mental structure (both informational and physical); a claim that it is not true for the case of instantiated concepts.

[35] In fact, the result of those cognitive processes is the only sort of empirical evidence that we have about what we call "concepts".

[36] Additionally, in the line of the conclusions drawn in section 5.3.1, I think that it cannot be said that (instantiated) concepts are something that we have, but merely that they are something in which we are competent.

INSTANTIATED CONCEPTS AND INFERENCES

But if instantiated concepts are mental events –in the sense of results of cognitive processes–, how could they uphold inferential processes? My answer is that they do it as far as the classification of an object *o* under a category *C* –as result of an instantiation process– can be an input of other subsequent cognitive processes. Or, in other words, instantiated concepts can support inferences because such inferences could work over the output of the former processes of categorization.

With regard to this, the temporal non-persistence of instantiated concepts –or, in Casasanto and Lupyan's terminology, the illusion that *ad hoc* concepts are stable over time– does not prevent that that those instantiated concepts may uphold inferences. The reason is that, in a framework like this, an inferential process is the result of two sequential stages:

(1) Classification of an object *o* under the category *C* –a process which results in the instantiation of the concept *C*, and a subsequent categorization of *o* under *C*–.
(2) Attribution to *o* of some of the properties of *C* about which the subject had no previous information for the case of *o*[37].

And, although these two stages are indispensable for an inference, only the second may be properly identified with an inferential process (i.e., with drawing a conclusion from a set of premises). Therefore, even though *ad hoc* concepts –or instantiated concepts– are unstable due to their dependence on context (and the fact that context varies from occasion to occasion); and even if that instability infected the classifications and inferences based in such instantiations[38], this does not prevent the existence of categorizations

[37] This inference/attribution of a property *P* (located in a dimension *f*, where the value of *o* is unknown for the subject) from a category *C* to an object *o* may be –at least– conceived in three different ways:

(a) It may be attributed the location of the prototype of *C* in the dimension *f*.
(b) It may be attributed the interval of values that the object *o* can take in the dimension *f*, based on both the actual instantiation of the concept *C*, and the location of *o*. How would that interval be determined? In the case of a concept with *n* relevant dimensions, and an object *o* with *m* known components (where $m<n$), for each unknown dimension *f* the object *o* is located, in a $(m+1)$-dimensional space, on the line L that (i) is orthogonal to the hyperplane determined by the axes of the *m* known dimensions, and that (ii) goes through the point representing *o* in that hyperplane. The interval of values that *o* can take in *f* is determined by the part of the line L contained by the region representing *C*. For example, in a case with 2 dimensions (x,y), and an object *o* with a known value $o_x=3$ and an unknown value o_y, the interval of possible values for o_y is the part of the line L, (i) orthogonal to the *x*-axis and (ii) that goes through the point $(o_x, 0)$, contained by the 2-dimensional surface representing *C*. Unfortunately, a process like this requires the determination of the region –or boundaries– associated to *C*, something that –as argued in section 5.2.2– is computationally very costly and, by virtue of this, psychologically implausible.
(c) It may be attributed, on the basis of the stored exemplars of that category *C* (see footnotes 27 and 33 in chapter 2), the mean –or median– of the *f*-components of the nearest exemplars to *o*. In contrast to (b), the computational cost of this alternative is much smaller.

[38] By virtue of this, a same object could be classified under various categories –depending on context–, and distinct properties could be attributed to it –in function of those different categorizations–.

standing on those instantiations and, subsequently, of inferences based on those classifications.

5.4. Storage and instantiation: There and back again

In this section I discuss the relation of mutual dependence that exists between the notions of stored concept and instantiated concept. Then I examine how is the process by virtue of which the instantiated concepts of a given category have an influence on its associated stored concept. However, before all this I will have to make explicit the kind of life cycle of a concept assumed in my work, and where I frame the aforementioned relations and processes.

5.4.1 Scheme of the life cycle of a concept

The structure of the life cycle of a concept –in a subject's mind– is a key element for the characterization of the existing relationships between the stored concept and the instantiated concept of a certain category. The aim of this part is to sketch the main stages of such life cycle, through which concepts are acquired, modified and applied in cognitive tasks:

(1) *Initial acquisition of a concept*: independently of the adopted approach, the first stage in any life cycle of a concept is always its initial acquisition (or formation). As a result of the acquisition process, our cognitive system stores certain information about that category. In particular, if the prototype theory is assumed –as done by me in this work–, concept acquisition takes place through a similarity maximization process that abstracts the underlying properties in a set of examples. Because of this, a prototype is produced for the new concept, whose location is registered in memory.

(2) *Subsequent revision* (or conceptual change): then, after the previous first acquisition of a concept, if the subject is exposed to new exemplars –or knowledge–, the information stored by the mind about that category may change as a result of conceptual readjustment processes. In the case of my view, the location of its associated prototype might be modified through a resemblance maximization process similar to that which gave rise to the original formation of the concept. Due to this, the information stored about that category will be updated with a new version of its associated prototype[39].

(3) *Application of a concept*: lastly, the information accumulated about each category will be used by the mind in cognitive tasks (i.e., categorizations, inferences, etc.) when that category is one of the relevant concepts in the considered context. In my view, the retrieved information about that category, together with informa-

[39] I am disposed to accept that a mental file may gather an historical of past versions of the information stored about a given category. Indeed, the fact that those historical versions were recorded under a same mental file is what would provide the continuity/stability required in order to explain our ability to accumulate new knowledge over time about categories.

tion about other concepts significant in that context, and other contextual parameters, will produce the application –or instantiation– of the concept in each particular occasion/circumstances.

It must be stressed that the previous general scheme is neutral, in the sense of valid for every empiricist view of concepts. That is, the three aforesaid stages are necessary phases of the life cycle of a concept, independently of the adopted assumptions on the nature and grade of contextual dependence of concepts –although I have described them for the particular case of a contextualist view of the prototype theory–.

5.4.2 *Mutual dependence between storage and instantiation*

And, even though my thesis is that storage and instantiation are two different notions –or senses– of concept, associated to distinct facets in their life cycle, I also think that there exists a strong mutual dependence between both notions:

— *Dependence of instantiation on storage*: as I have explained throughout this entire chapter, when a concept C is used in cognitive tasks such as categorizations, inferences, etc., some of the information stored about C (i.e., part of its stored concept) is read from memory and applied –together with data about other relevant categories and other contextual factors–.

Consequently, the instantiated concept depends on the stored concept because part of the information accumulated relative to that category is used by the cognitive processes which instantiate such a concept.

— *Dependence of storage on instantiation*: as said in the previous section, as a result of the initial acquisition of a concept C some information is stored about it. In my view, that information –or *stored concept*, in my terminology– is the location of the prototype associated to C –call it P_C–. The stored concept is maintained in a stable and persistent way until new information triggers a revision of it. As a result of that revision, the location of the new prototype P_C^* may be different and, in such a case, P_C^* will be registered as a new version of the stored concept associated to that category.

However, the important point here is that any subsequent revision which produced a new prototype P_C^* (i.e., a new version of the stored concept) would be the result of a maximization process whose input might contain objects recently classified under that category by the subject –in past applications of that concept–. The idea is that the classification of a new object under a category results in a new exemplar that have to be considered by the learning processes[40] that determine the location of the prototype associated to C. But, because the members of that category have changed since the last execution of the learning processes –either in the original acquisition of the concept or in a future recalibration of it–, a new location of the prototype could be produced. As a consequence, the information rec-

[40] For a description of how those processes work, see section 4.1 in this doctoral thesis.

> orded in the mental file of that category has to be amended with a new (version of that) stored concept[41].
>
> Therefore, since former categorizations of objects as *C* –resulting from past instantiations of the concept *C*– may be used by subsequent readjustment processes that will modify its associated stored concept, it is fair to say that stored concepts can, and usually do, depend on instantiated concepts.

Having said this, it should be clear that the life cycle of a concept described in section 5.4.1 is not linear, but circular. There the registered information remains stable –in the stored concept– until its next occasion of use, when the concept is instantiated again. However, after new instantiations of that concept, and as a result of the classification of new objects under that category, its associated prototype would be updated through a process of conceptual readjustment. Consequently, after the initial acquisition of the concept, its life cycle consists in a loop between successive revisions and applications –or instantiations– of that concept.

5.4.3 On the process of conceptual readjustment

Throughout this chapter it has been explained how instantiated concepts can be produced from the information stored by the mind about their associated categories, by virtue of which I have held that instantiations depend on storage. However, little has been said about the processes of conceptual readjustment owing to which the stored concepts also depend on previous instantiations of their respective categories. Because of this, the aim of this section is to account for how the instantiated concepts have an influence on the information registered in memory about categories.

To take a toy example, let me consider the case of two categories *A* and *B* whose associated prototypes P_A and P_B were initially acquired[42] on the basis of the two sets of exemplars $\mathbf{I}_A$ and $\mathbf{I}_B$, whose members are represented in Fig. 5.6a by means of ×'s. The locations of those prototypes are determined through the following formula:

$$P_C = \frac{\sum_{e \in \mathbf{I}_C} e}{\sum_{e \in \mathbf{I}_C} 1}$$

where *e* represents an exemplar belonging to the set $\mathbf{I}_C$ of members of the category *C*.

As a consequence, the conceptual space is partitioned into two regions, which are delimited by the boundary also shown in Fig. 5.6a. After that, new objects may be categorized under *A* –or under *B*–, like those represented in Fig. 5.6b by means of +'s, as a result of later instantiations of *A* –or of *B*–. Because of this, it is possible to say that the set $\mathbf{E}_A$ of new examples categorized as an *A* is a function of past instantiations of the concept *A*; and the same may be said regarding the dependence of the set of new exemplars $\mathbf{E}_B$ on past instantiations of *B*.

[41] For more on this, see section 5.4.3 below.

[42] Under the assumption of a Euclidean metric where all concepts and dimensions are equally weighted.

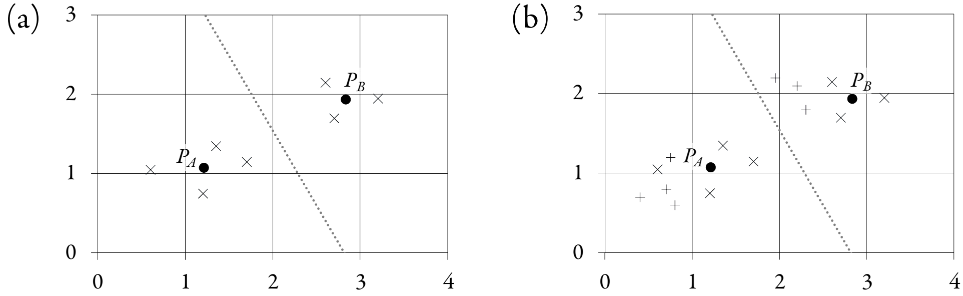

Fig. 5.6. Example of conceptual change (I): original acquisition and classification of exemplars, for a Euclidean metric where all the concepts and dimensions were equally weighted. (a) Original acquisition of the concepts *A* and *B*, from two sets of exemplars $\mathbf{I}_A$={(0.6,1.05),(1.2,0.75),(1.35,1.35), (1.7,1.15)} and $\mathbf{I}_B$={(2.6,2.15),(2.7,1.7),(3.2,1.95)}, whose locations are represented by ×'s. The resulting prototypes of those categories, P_A=(1.213,1.075) and P_B=(2.833,1.933), are represented by means of black dots; and a grey dotted line represents the boundary between their associated regions. (b) Categorization of new exemplars –represented through +'s– under the category with the closest prototype. In particular, the examples in $\mathbf{E}_A$={(0.4,0.7),(0.7,0.8),(0.75,1.2),(0.8,0.6)} are classified as cases of *A*, while the ones in $\mathbf{E}_B$={(1.95,2.2),(2.2,2.1),(2.3,1.8)} are categorized as cases of *B*.

At this point, the sets of exemplars of *A* and *B* known by the subject (i.e., $\mathbf{I}_A \cup \mathbf{E}_A$ and $\mathbf{I}_B \cup \mathbf{E}_B$, respectively) are greater than the original ones (i.e., $\mathbf{I}_A$ and $\mathbf{I}_B$). Taking all this into account, and since a subsequent revision of the concept *A* –or of the concept *B*– will be based on the extended set $\mathbf{I}_A \cup \mathbf{E}_A$ –or $\mathbf{I}_B \cup \mathbf{E}_B$, for the case of *B*–, that later revision will usually produce a prototype P_A^* distinct from P_A –or a prototype P_B^* distinct from P_B in the case of *B*–. Those new prototypes P_A^* and P_B^* are represented in Fig. 5.7a by grey dots, whose locations may be determined as follows:

$$P_C^* = \frac{\sum_{e \in (\mathbf{I}_C \cup \mathbf{E}_C)} e}{\sum_{e \in (\mathbf{I}_C \cup \mathbf{E}_C)} 1}$$

Consequently, the boundary which delimits the regions associated to *A* and *B* (which is represented in Fig. 5.7a through a grey dashed-and-dotted line) will be different from the one produced by the initial prototypes. As a result, cases previously categorized as *A* (e.g., o_1 and o_2, represented in Fig. 5.7b by triangles) are now classified under the category *B*.

Finally, the locations of the new prototypes (i.e., P_A^* and P_B^*) will be registered as new versions of the stored concepts associated to the categories *A* and *B*. And, because their locations are a function of the instantiations of *A* and *B* for the objects in $\mathbf{E}_A$ and $\mathbf{E}_B$, respectively, it is fair to say that stored concepts depend on instantiated concepts.

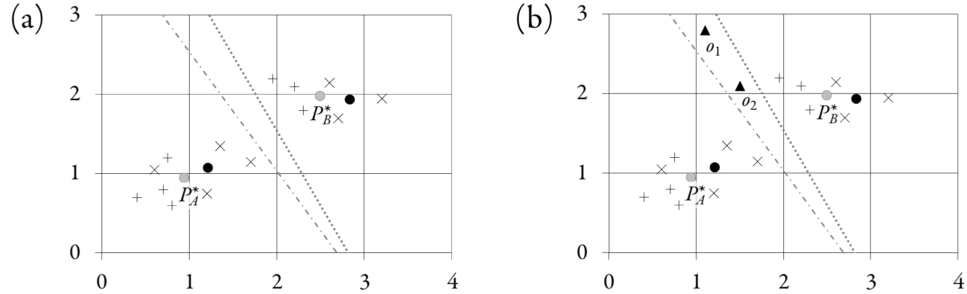

Fig. 5.7. Example of conceptual change (II): new prototypes and reclassification of exemplars, for a Euclidean metric where all the concepts and dimensions were equally weighted. (a) Subsequent revision of concepts *A* and *B*, from the extended sets of exemplars $I_A \cup E_A$ and $I_B \cup E_B$. The new prototypes of those categories, P_A^* =(0.938,0.95) and P_B^* =(2.492,1.983), are represented by grey dots; and a grey dashed-and-dotted line represents the new boundary between the conceptual regions of *A* and *B*. (b) Categorization of the exemplars o_1=(1.1,2.8) and o_2=(1.5,2.1) –represented by means of black triangles–, on the basis of the new prototypes P_A^* and P_B^*, under the concept *B*; in contrast with their classification as *A*, on the basis of the original prototypes P_A and P_B.

5.5. *Intermediate discussion*

In this section I will consider some issues related with the topic of this chapter. In the first place, I compare my approach with other leading perspectives (for instance, those of Machery, Casasanto & Lupyan, Prinz and Barsalou), in order to clarify as neatly as possible where my view and conclusions agree and disagree with the ones of these authors. Next, I highlight the major strong points of my proposal, for example, its ability to combine invariantist virtues within a contextualist framework, and the absence of a conceptual core –or membership conditions– associated to categories.

5.5.1 *Comparison with similar views*

In the previous sections of this chapter I have occasionally compared my point of view in this thesis with other prominent positions in the field, with the aim of delimiting how my perspective differs from those other approaches.

For instance, I have argued that my stored concepts should not be identified with Machery's (invariantist) notion of concept –understood the latter as bodies of knowledge in the long term memory that are used by default by our cognitive processes (see section 5.3.3)–. In particular, Machery's view has the problem of postulating the existence both of *default* bodies of knowledge and of a *neutral* context which works by default in those kinds of cognitive tasks. Unfortunately, there is no evidence of the existence of *default* bodies of knowledge used in categorizations, nor of the existence of a *neutral* context in all those cognitive processes (see footnote 30 in this chapter).

Anyhow, more interesting is the comparison with other views, neither in complete opposition nor in whole alignment, but which share a –more or less– family resemblance with my perspective. That is the case of Barsalou's (1987) and Prinz's (2002)[43]. In particular, these latter –together with Casasanto and Lupyan's– are related proposals, with elements in common, so it will be useful to contrast my view with those of these authors[44]:

— On the one hand, according to Prinz a concept consists in a set of mental representations (called by him *proxytypes*) stored in long term memory, that are activated –or copied– in working memory for representing a given category. In this approach, the context determines which proxytype (from among all the proxytypes associated to that concept) is activated in each particular occasion.

  Firstly, my view resembles Prinz's in that both approaches distinguish the storage of information (associated to a concept) in long term memory, from the use of that information –or of part of it– by the working memory in tasks such as categorization –and other similar ones– in a context-dependent manner.

  Nevertheless, two key differences distinguish both proposals: (1) Regarding the information assumed to be stored in long term memory, Prinz's proxytypes are different versions of the same concept, each of them specific of a particular context. By contrast, I do not argue for the existence of various stored concepts (i.e., one for each context), but for only one unique stored concept for each category, which will be differently instantiated depending on context. (2) With regard to how the stored information is applied in cognitive tasks such as categorization, Prinz' activation of proxytypes is a mere copy from long term memory to working memory of the information recorded under a given proxytype. In contrast, my view proposes a process of instantiation –not of copy– of that information which, together with other context-dependent factors, will produce what I have called the instantiated concept.

  Thence, my proposal –in contrast to Prinz's– avoids the difficulties due to the assumption that a different version of each concept (i.e., a different proxytype, in Prinz's terminology) is stored for each possible context, and the subsequent multiplication –potentially unlimited– of the number of stored proxytypes.

— On the other hand, my approach is quite close to Barsalou's, as long as his graded structure of concepts is the result of processes of similarity comparison (Barsalou 1983; 1987). In respect to this, the main difference is that while Barsalou (1983, p. 212) seems to conceive similarity in line with Tversky's feature model (1977), my view is more general, since it is based on dimensional models[45].

[43] There are other discriminations in the literature that, although on the surface they resemble my distinction between stored concepts and instantiated concepts, in reality they are much less close to my view than Prinz's or Barsalou's positions. For example, regarding Carey's (2009; 2011) distinction between concept and conception, and even though his idea of *concept* –as that which determines the meaning– might be charitably identified with my *instantiated concepts*; his notion of *conception* –as that we believe about categories– can be no way identified with my *stored concepts*.

[44] Since the similarities and differences between my view and that of Casasanto and Lupyan have already been elucidated in section 5.2, they will not be discussed again in the present section.

[45] And recall that –as said in section 4.1– the feature models are a particular case of the dimensional ones.

Barsalou's main point (1987; 1993) is that concepts are not invariant structures stored in long term memory –and waiting to be intactly retrieved or copied in working memory when needed–; but provisional constructs produced in working memory (specifically for each occasion), taking as input information stored in long term memory. In respect to this, my proposal is completely equivalent to Barsalou's.

Nonetheless, whilst Barsalou (1987, p. 114) does not accept that our mind can contain invariant cognitive structures associated to categories[46]; the existence of stored information that remains stable between consecutive reviews of a certain concept constitutes the cornerstone of my stored concepts. Another difference is that Barsalou (*ib*., pp. 116) only includes within information relevant for the construction of a concept in working memory, information which provides expectations about that category, but he does not include information associated with other relevant concepts in that context –something that has a key role in my proposal–.

Lastly, when Barsalou claims that "the same concept is rarely if ever constructed for a category" (Barsalou 1987, p. 101), the idea reminds us of my instantiated concepts. However, my notion of *instantiated concept* does not seem to be present in his work, as evidenced when he introduces the idea of *concept* as follows:

> [C]oncept will refer to the particular information used to represent a category (or exemplar) on a particular occasion. (Barsalou 1987, p. 116)

That is, according to Barsalou *concepts* are the instantaneous inner information –or, in Clark's (1993, p. 93) terminology, *occurrent states*– that determines the categorization performance in a particular context. By contrast, on my view instantiated concepts are the result of applying that information, not such information *per se*. Thus, the main difference is that Barsalou (1987; 1989) seems to attribute an ontologically strong character to the construals produced in working memory (as something that is, in fact, there and over which the mind can operate), and that is not the way I think about instantiated concepts[47].

5.5.2 Advantages of this approach

One major advantage of my approach is that, although it characterizes a strong version of contextualism –as it is the *ad hoc* cognition framework–, it is likewise able to explain, from that contextualist standpoint, a phenomenon typically identified with the invariantist perspective. Indeed, on the one hand the proposed model articulates a contextualist framework compatible with the evidence against the existence of definitions –or

[46] Anyhow, Barsalou is not completely clear regarding this issue. In fact, sometimes he claims that there exists relatively stable knowledge about categories in long term memory (Barsalou 1989), while in other cases he rejects the existence of a set of invariant structures used recurrently by the mind (Barsalou 1987; 1989).

[47] See above for the difference between concepts as *mental entities* and concepts as *mental events*.

conceptual cores–, and thus able to provide an account of our adaptive abilities to changing environments. But, on the other hand, my proposal is also able to explain how, although instantiated concepts are absolutely context-dependent, stored concepts are stable enough to be the means by which new information about concepts can be collected over time.

Secondly, there is nothing in my approach that could be identified with a conceptual core (nor, as a result, with the conventional meaning of a concept –or term–), so it is meaningless to ask for the conditions of application of a certain concept. In fact, since all there is in my proposal are mental processes classifying objects into categories –in an absolutely occasion-specific way–, the model works perfectly well without having to resort to the idea of definition (or to a stable set of membership conditions).

Finally, the presumed theory on the structure of concepts and its cognitive deployment (i.e., a prototype theory characterized by means of a geometric similarity space, where different versions of the prototype of a concept may be stored within the same mental file) allows to explain, not only how concepts are learned/acquired and, afterwards, are reviewed/modified; but also how distinct (previous) variants of the same concept may be present in the mind.

5.6. Instantiated concepts, persistence, and representationality

Throughout this chapter, I have claimed that only stored concepts can be conceived in terms of persistent entities, even though those responsible of the external manifestation of categories are the instantiated concepts. I have also held that instantiated concepts should be thought to be mental events –dependent on the actual context– that occur at the end of the cognitive processes (i.e., categorizations, inferences, etc.) which display them, and that, in consequence, they have no minimal persistence. The aim of the present section is to clarify this latter thesis –namely, that instantiated concepts are not persistent–, and to analyze its implications for the representational (or non-representational) character of that notion of concept.

As said in chapter 1 regarding the question of representationality, one dominant strand in cognitive science considers that concepts are representations, that is, mental particulars with semantic properties. That is the view of classical computationalism, which –rooted on the Representational Theory of the Mind– presumes that psychological states and processes occur within an inner representational system. If, as said by Fodor "there is no computation without representation" (Fodor 1981b) and, if concepts are conceived within a computational framework –as I have done in my doctoral thesis–, then concepts would have to be representations since, otherwise, the calculations required by the cognitive processes involving those concepts (e.g., by categorizations, inferences, etc.) are not possible.

However, since the advent of connectionism the notion of mental representation has been called into question by many supporters of this perspective[48], who maintain that in

[48] More recently, mental representations have also been questioned by advocates of enactivism and embodied cognition.

most neural networks semantic properties cannot be attributed to any individual element (i.e., unit or node) of the network, so representations could not be located beyond the input layer of the neural network. The point is that, under this last view, concepts cannot be mental representations.

By contrast, my thesis in this section is that the key reason to reject that concepts are mental representations is not the connectionist –or classicist– architecture of the brain, but the *minimal persistence* condition demanded of representations, which cannot be satisfied in a contextualist view of the mind. Thence, I will show that, under the assumption of contextualism, instantiated concepts lack minimal persistence and, as a consequence, that they cannot be a representation of their associated categories.

5.6.1 The issue of persistence

In section 5.3.3 I suggested that instantiated concepts might be thought to be mental events that happen at the end of the cognitive processes which instantiate them in each particular context. The aim of the present section is to clarify, and argue in favor of, the thesis that instantiated concepts are non-persistent mental events.

My point is that instantiated concepts should be viewed as the result of cognitive processes, not in the sense of *products* –i.e., minimally persistent psychological entities stored in mental structures–, but in the sense of *phenomena* –as that which happens when something is categorized under a given category–. Thus, they should be identified, not with stable mental states or entities, but with punctual psychological events.

The product/phenomenon distinction

Here I contend that processes –whether cognitive/computational or not– may culminate in two types of result, namely, products and phenomena. (I will also suggest that the product/phenomenon distinction is graded, since the boundary between products and phenomena depends on the subject's perspective and interests, and –more particularly– on their temporal horizon.) On the one hand, products would be minimally persistent results, which can typically be accessed in a future time[49]. On the other, phenomena are, conversely, non-persistent results, which cannot be accessed beyond their occurrence time.

However, the product/phenomenon distinction becomes clearer when analyzing examples outside the field of cognition. Regarding the notion of *product*, let me examine the case of an algorithm that calculated the square root of a number n, and stored the result within a memory register at time t_p, when the computations end (see Fig. 5.8). In this case, the result of the process can be identified with the physical state of a set of transistors b_1, b_2, b_3, ..., which codify the outcome of the square root algorithm. Such a result can be called a product because the logical states of b_1, b_2, b_3, ..., can be accessed beyond

[49] Although in general products are minimally persistent, they may sometimes be punctual results. For instance, this would be the case of the non-minimally-persistent intermediate results in a computer system. For my part, even though I do not exclude that possibility, such cases are not relevant for my discussion. [Thanks to Robert Goldstone for highlighting this point at the Workshop on Concepts in Action: Representation, Learning and Application (Osnabrück University, August 2018).]

the instant t_p when the computation process ended. Or, in other words, since those logical states are persistent, they can be subsequently accessed by other processes different from the one that produced them.

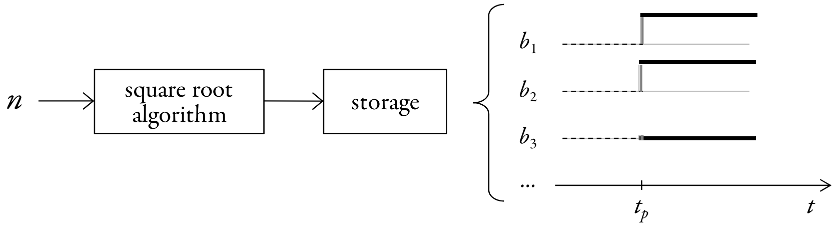

Fig. 5.8. Example of a process whose result is a product: algorithmic computation and storage.

By contrast, with respect to the notion of *phenomenon*, a controller system which periodically received a temperature level from a thermal sensor, and then emitted a flash of light when a threshold is surpassed (see Fig. 5.9), would be a process whose results are phenomena. The result of this process can be either the emission state e_0 –if the thermal threshold has not been exceeded, and no emission occurs–, or the emission state e_1 –if temperature is above the threshold, and a light/sound signal is emitted–. In this second case, the result of the considered process is a non-persistent phenomenon –for example, a millisecond photon emission–, which happens at time t_p, and cannot be accessed a couple of seconds after the end of the process. That is, it is not possible to consult the result of this process after its occurrence time t_p.

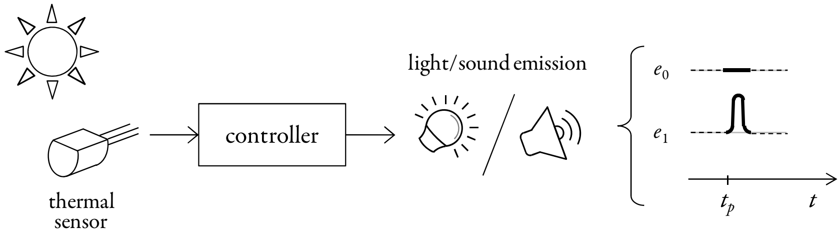

Fig. 5.9. Example of a process whose result is a phenomenon: thermal control and alarm.

This distinction –although not expressed in terms of products and phenomena– is also present in the inner workings of a computer. There we find some states which may be called persistent, for instance, the information stored in hard disks, in random-access memory, or in cache memory; while other states should be called non-persistent, like program memory and, to a lesser extent, processor registers (i.e. CPU instruction registers). Thence, even though we could speak of persistent states when referring to data stored in registers of the first kind, that is not possible for the state of a CPU, which could not be called "persistent" because the program memory is continuously changing.

Nonetheless, the product/phenomenon distinction is a graded one, given that the boundary between products and phenomena is vague and depends on the subject's perspective and interests. In this sense, there is no product –resulting from a process– absolutely stable, and there is no phenomenon absolutely non-persistent. The reason is simple: no product has infinite duration, and no phenomenon has zero duration. The idea is that this distinction is always relative to a concrete temporal horizon and, for the particular case of computational –and cognitive– systems, is relative to their clock rates. For example, if the duration of the inner result *R* of a computer device is not greater than twice, thrice, or a few more times its clock frequency, then such a result will not be available beyond those processing cycles. Because of this, *R* should be called a *phenomenon*, given that it is non-accessible to other processes different from the one which produced *R* –and those immediately subsequent to the generating process–.

The graded character of this distinction is also evident when we look at the result of processes in nature. For instance, rain, lightning, mist, frost, storms, hurricanes, and so on, are all called (weather –or meteorological–) phenomena; earthquakes and volcanoes are also called (geological) *phenomena*. By contrast, mountains, caves, faults, etc., are called (geological) formations/structures –or, in my terminology, *products* of geological processes–, because they are invariant in our lifetime. Nonetheless, the results of other geological processes (i.e. periodic displacements of orogenic regions or formation of metamorphic minerals), when evaluated in geologic time scale, can be called *phenomena*.

Finally, a peculiar feature of computational and cognitive processes is the fact that their inner states are only accessible to its own computational/cognitive system. That is the reason why in this sort of processes the product/phenomenon character of their results must be determined according to the temporal horizon of the own system. Or, in other words, the own computational/cognitive system is the only possible reference to determine whether the results of its processes are either products or phenomena.

INSTANTIATED CONCEPTS ARE PHENOMENA

Here my thesis is that instantiated concepts are the result of mental processes of the second kind –i.e., they are *phenomena*–, which build them for each occasion-specific context. At this point, it is worth comparing the temperature controller example with the case of instantiated concepts, in order to show why the latter should be also called *phenomena*:

- — *Both are context-dependent processes*: On the one hand, the thermal threshold can depend on season, on the moving average of *n* past temperatures, etc., which act as the context of this control system. On the other, instantiation processes can be influenced by the relevant properties –what may depend on circumstances–, by the situation –i.e., concepts considered in a near time interval–, etc., which would play the role of the context of instantiation.
- — *Both are occasion-specific construed*: Regarding the controller system, light/sound alarm signals are specifically produced in each particular occasion, depending on the current thermal threshold. In the case of concepts, instantiated concepts are specifically created on each occasion of use, in function of the current relevant properties –and other contextual factors–.
- — *Both are one-time evaluated*: In the first case external temperature might be evaluated according to a schedule –or on demand– and in such a case the controller would not be continuously working. Analogously, concepts are instantiated only

when subjects categorize, draw inferences, communicate, etc., so instantiation processes are not continuously running.

All in all, the flash of light/sound emitted by the thermal controller is a non-persistent *phenomenon* because it is produced in a context-dependent way and its duration is less than the temporal horizon of the alarm system (i.e., the inverse of its schedule frequency). By virtue of this, it can be said that the flash signal resulting from this system only exists when the controlling process ends, and its emission happens. Almost the same may be said regarding instantiate concepts. They are produced in a context-dependent way by instantiation processes (e.g., in categorization tasks), and their duration is less than the temporal horizon of the cognitive system, because after the end of the instantiation process context may change and, in consequence, the result of any other subsequent instantiation process could be different. Thus, it can be said that the instantiated concept resulting from a categorization task is a non-persistent / punctual *phenomenon*[50], susceptible of being identified with a mental event, which only exists at the moment when its instantiation process ends.

To sum up the discussion regarding the persistence of instantiated concepts, and under the non-problematic assumption that (instantiated) concepts are the result of mental processes, first I have shown that the result of a process can be either a product or a phenomenon, and then I have argued that instantiated concepts are phenomena. On this basis, and given the non-persistent / punctual character of phenomena, I have concluded that *instantiated concepts are non-persistent / punctual* mental events that occur in the mind.

5.6.2 The issue of representationality

Now I will argue that if instantiated concepts lack minimal persistence, as I have held in the prior section, then they cannot be considered a representation of their respective categories.

Representational and non-representational approaches to concepts

As said in chapter 1, one dominant view in psychology and cognitive science is that concepts are mental representations, that is, particulars with semantic properties (i.e., truth values/conditions, satisfaction values/conditions, reference, content, etc.). This approach is rooted on the Representational Theory of the Mind (RTM), according to which all the different kinds of psychological states and processes occur in an internal representational system (i.e., any kind of thinking involve mental representations). One popular

[50] My use of the term "phenomenon" should not be confused with other similar ones like, for instance, the expressions "occurrent state" and "instantaneous internal state" used by Clark (1993, p. 93) in order to describe Barsalou's sense of concept, as "the particular information used to represent a category (...) on a particular occasion" (Barsalou 1987, p. 116). However, even though I sympathize with both expressions –because they might describe the phenomenal character of instantiated concepts–, my view of them is very different from Barsalou's one, so the differences are not merely terminological. The reason is that Barsalou seems to think that those concepts represent categories, but that is precisely the opposite of my argument's conclusion.

way of characterizing RTM is in terms of a language-like syntax and a compositional semantics, just like Fodor[51] (1987) does with his view of RTM as a symbol system formed by mental representations where (i) propositional attitudes are conceived as bearing computational relations to mental representations; and (ii) mental processes are causal sequences of tokenings of those mental representations –or, in other words, causal interactions among representations–.

Surely the foremost contemporary version of RTM is the Computational Theory of Mind (CTM). On this view, the *mind* is conceived as a sort of computer; *mental processes* as computations; and *mental states* as results of the occurrence, transformation and storage of informational structures (Pitt 2017). Under this approach *mental states* can again be seen as computational relations to mental representations, and *mental processes* as sequences of those mental states. However, there is a significant disagreement among advocates of CTM regarding the implementation of cognitive states and processes, in particular about whether it should be classical or connectionist. On the one hand, the proponents of classical architectures hold that mental states –or mental representations– are symbolic structures constituted by semantically evaluable objects, and that mental processes are operations ruled by sequences of those representations (Turing 1950; Fodor 1975, 2000, 2008; Newell and Simon 1976; Marr 1982). On the other, the defenders of connectionist architectures hold that mental representations are carried out by patterns of activation in a network of nodes –or units–, and that mental processes consist in the formation and spreading of those patterns of activation (McCulloch and Pitts 1943; Smolensky 1988; Rumelhart 1989).

Nowadays RTM remains the primary view in many areas of cognitive science. It is the core of a strong and influential model of thinking, mental states and processes, and the position against which many others discuss (Margolis and Laurence 2007; Ramsey 2017; Rowlands 2017). Even some advocates of connectionism seem either to accept the idea of representation (McClelland and Rumelhart 1985; Smolensky 1988, 1991; Hinton 1989, 1990; Elman 1990b, 1991) –sometimes under the label of *distributed*, in the sense of *non-localist*, representations[52]–, or to be reluctant to its complete repudiation (Clark

[51] Fodor's view of RTM is personified in his *language of thought hypothesis* (Fodor 1975), to the point that the inner representational system presumed by RTM could be identified with the language of thought.

[52] When connectionists describe the notion of *representation* in terms of distributed codification of information, they usually say that entities are represented by patterns of activity distributed over multiple cognitive/computing elements –i.e., nodes or units– (Hinton, McClelland and Rumelhart 1986, p. 77; Elman 1990b, p. 351; 1991, p. 210); and that the network's ability to meet a goal condition –or to solve a particular task– in a particular environmental situation allows to identify the network's associated internal state with a veridical representation of the corresponding environmental state (Smolensky 1988, p. 15; Elman 1991, p. 195).

However, such a view of representation is controversial because, even if we accept that a pattern of activity able to identify a particular *entity* constitutes a representation of that entity, it remains the question of whether those patterns may be considered a representation for the case of *categories*. Regarding this last point it could be argued that a pattern of activity cannot be a representation of a whole category due to two main reasons:

1993; Clark and Toribio 1994). By contrast, other connectionist approaches to CTM maintain that there are no mental particulars susceptible to be identified with representations[53] (Ramsey, Stich and Garon 1990; Ramsey 1997, 2017). Under this last view, nodes –or units– are not taken to be semantically evaluable and, consequently, informational structures are not conceived as constituted by semantically evaluable objects. Therefore, the defenders of this perspective hold that it should not be spoken of mental representations, but of distributed codification of information by means of levels, weights and nodes.

My view is closer to these last connectionists –according to whom concepts are not mental representations–. Notwithstanding this common endpoint, in the rest of this section I will try to show that this very same conclusion can be reached taking as starting point an assumption, not about the computational/neuronal architecture of the mind –as those connectionists do–, but about the degree of contextual dependence of concepts. In fact, my thesis is that the key reason to reject that (instantiated) concepts are representations is not the connectionist or classicist architecture of the mind, but the *minimal persistence condition* demanded of representations –which, as said above, is not satisfied by the sense of concepts as instantiation–.

On the notion of representation

Concerning the notions of representation –in general– and mental representation –in particular–, it is important to observe that the idea of *mental representation* is not a commonplace in ordinary discourse, but an assumption within a theoretical framework. Therefore, as noticed by Cummins (1989), it is naïve to expect that such a notion is the same for classic computationalism, connectionism, folk psychology, neuroscience, and many others who make use of such an idea. Nevertheless, my aim in this section is not to provide a definition of it common to all those fields, but to show that –independently of the considered framework– *minimal persistence* is a condition tacitly present in the majority –or maybe even all– of them[54]. Thence, it does not matter how mental representations are conceived (i.e., as something localized or distributed; continuous or discrete; as a

(1) If patterns of activity are context-dependent, then there will not be a unique pattern for each category, but a set of patterns –one for each particular context–, so none of them could wholly represent that category.

(2) A pattern of activity working as a classifier –i.e., able to determine whether a concrete entity meets a set of conditions– evaluates, not the whole universe of possible entities, but only the particular entity *a* considered. Thence, that pattern will be –at best– a representation of the entity *a*, but not of its entire associated category, since the pattern of activity applied to identify another entity *b*, belonging to the same category, might be different from that required for the identification of *a*.

[53] Indeed, the discussion between representationalists and anti-representationalists is the core of a promising and fruitful debate, as evidenced by the abundant literature on this issue (Stich 1983, 1992; Brooks 1991; Clark and Toribio 1994; Sutton 2004; Margolis and Laurence 2007; Ryder 2009; Egan 2014; Ramsey 2017; Rowlands 2017).

[54] Given that there are few who explicitly recognize the role played by persistence as a required condition of any representation (Shagrir 2012; Danks 2014).

binary or ternary relation; or in causal, informational or teleosemantic terms), because the only thing which is required by my argument is the attribution of a minimal persistence to whatever is called "mental representation".

As said above, possibly the most prominent view in cognitive science is that according to which mental representations are internal psychological states with semantic properties. Nevertheless, there are very different ways in which those mental states might be conceived. They can be thought –for instance– as continuous or discrete, as distributed or local, etc. (Pitt 2017). This notwithstanding, a minimal persistence seems to be presumed in all these cases. Indeed, a good approach is to discuss one by one, first those different alternatives, and then the other distinct ways in which the idea of representation may be understood. In that way, regarding the distinction between continuous/analog and discrete/digital representations (Goodman 1968; Lewis 1971; Pylyshyn 1980; Dretske 1981; Haugeland 1981), nothing in that distinction suggests that the underlying states are non-persistent[55]. This is particularly true when cognition is conceived as governed by rules acting on a symbolic system or scheme, regardless of whether representation is deployed in engineered or biological terms (Eliasmith and Anderson 2003).

A more interesting distinction is the one between localized and distributed representations, which emerge in the debate between connectionists and classical computationalists[56]. According to the classical view, mental representations are symbolic structures ruled by operations sensitive to their constituents (Turing 1950; Newell 1980; Marr 1982; Fodor and Pylyshyn 1988). This approach may be summed up as follows:

> Representation is simply another term to refer to a structure that designates:
> X *represents* Y if X designates aspects of Y, i.e., if there exist symbol processes that can take X as input and behave as if they had access to some aspects of Y. (Newell 1980, p. 176)

> A representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. (Marr 1982, p. 20)

Having said this, all of the elements in this perspective demand a minimal persistence. On the one hand, mental representations are *localized* in representational structures, schemes or states available to be invoked –or manipulated– in computations by those operations that involve them. On the other, mental processes are *specified* as information-processing machines that determine the inner workings of the system. Thence, it seems that classical computationalism tacitly assumes that mental representations are minimally *persistent* structures –able to encode properties about their represented objects– ready to be accessed by the cognitive processes working on them.

By contrast, connectionists maintain that mental representations are carried out by patterns of activation in a neural network, and that mental processes consist in the for-

[55] This is so independently of whether the pairs of notions continuous-analog and discrete-digital are distinguished or not (Maley 2011).

[56] Concerning this, *classicists* consider that representations are local –with regard to computation–, while *connectionists* hold that representations are distributed across large populations of units –none of them corresponding to the represented objects alone– (Thorpe 2003).

mation and spreading of those patterns (Smolensky 1988; Rumelhart 1989). In this case it is said that mental representations are distributed because they are not located in a particular node or unit, but spread across large sets of nodes. However, this is precisely the reason why many argue that connectionist networks have no mental objects that might be identified with representations; i.e., because there are no inner content-bearing particulars in the network that can play the role of representations (Ramsey, Stich and Garon 1990; Ramsey 1997)[57]. Against this criticism, it is often claimed that such patterns of activation can be described as states of the system –encoding either concepts or properties– (Fodor and Pylyshyn 1988; Rogers and McClelland 2004). According to this, there would be a sense in which a pattern of activation may "represent" a concept when that pattern classifies something under a category. Here my point is that, under connectionist representationalism, although representations were conceived as distributed patterns of activation, those patterns of activation are a kind of distributed coding of information which remains stable across time (Thorpe 2003)[58], so minimal persistence is also a condition required for them.

We could continue reviewing other different approaches to the notion of representation –i.e., causal theories, informational theories, historical/teleosemantic perspectives, etc. (Rowlands 2017)–. However, on my view such an effort would be unnecessary at this point, given that all these and other forms of contemporary representationalism typically conceive representations as neural states that stand in an appropriate relation to something else[59] (Fodor 1980b; Millikan 1984; van Gelder 1995) and, again, it seems that the assumption of a minimal temporal persistence is a condition required for all those states to be a representation.

Argument against the representational character of instantiated concepts

At this point, if representations can be described as relatively stable and persistent objects that codify information –because, as argued above, it seems that minimal persistence is a condition demanded of representations by any cognitive theory–, then my conclusion is that instantiated concepts cannot be representations since they do not encode stable information.

Therefore, the main reason to reject the presumed representational character of instantiated concepts would be the minimal persistence requirement demanded of the notion of representation (which is not fulfilled in a contextualist view of concepts), and not the structure/architecture –either classicist or connectionist– of the mind.

[57] Sometimes it is argued that these arguments against the notion of representation in connectionism constitute a challenge for a much wider range of approaches (Bechtel 1998).

[58] In the same line, Rogers and McClelland (2004, p. 50) remind us that one of the strong points of connectionist networks –demonstrated by Hinton (1981; 1986)– is their ability to store/represent a proposition by means of a stable pattern of activity.

[59] Those states might be called either a *representation*, or a *detector/classifier*, or an *indicator*, of the considered external object (Ramsey 2007; Ryder 2009).

5.7. Generalized argument

In this chapter I have focused my discussion in the approach chosen by me in order to characterize the idea of concept, namely, a geometric similarity-based framework articulated by means of the prototype theory of concepts. In spite of this, none of the key elements in my argument for the need to differentiate two distinct notions of concept crucially depends on the selected framework. Based on that, in this last section I will briefly sketch how such an argument may be generalized to any position that assumes a contextualist view of concepts.

In the first place, remember the main thesis of radical contextualism, namely, that *concepts have to be instantiated for each particular context*. This proposition is the main premise of my argument; a premise that can be broken down in three subsidiary theses –all of them more or less explicitly accepted by most advocates of contextualism–:

(1) Concepts have to be instantiated for each particular occasion.
(2) The instantiation of a concept depends on the considered context.
(3) Since those instantiations depend on context, they can lead to different external manifestations (e.g., categorizations) in function of the relevant context.

Therefore, for any contextualist approach to concepts, there is a sense of concept that (a) it is responsible of the external manifestation of each category, and (b) it may vary from one occasion –or context– to another. This sense of concept would correspond with my notion of *instantiated concept* –or *concept as instantiation*–.

Having said this, if the notion of instantiated concept is the result of a modulation –or instantiation– specific of context, thence it ought to exist something to be modulated. Indeed, for the case discussed in this chapter (i.e., a similarity-based framework characterized through a prototype theory), the thing that is modulated are the locations of the prototypes of the relevant concepts in each context. That is what I have called *concepts as storage* –or *stored concepts*–.

Nevertheless, this second notion of concept is not exclusive of prototype-based approaches since, as argued above, in any contextualist view there must exist something that is modulated by the instantiation processes[60]. Indeed, it is possible to test this assertion if we examine some other theories on the structure of concepts, like the classical and the exemplar views. Firstly, the case of the *exemplar theory* of concepts is very similar to the prototype-based view analyzed up to this point of the chapter. The idea is that, since the similarities between prototypes/exemplars and objects are determined through the very same computational scheme (namely that described in section 5.2.3) both for the prototype theory and for the exemplar theory, both of them will depend on the same contextual factors (i.e., relevant concepts/exemplars, type of metric, importance of dimensions, and weighting of concepts/exemplars[61]). On this basis, that which is modulated –in the

[60] Certainly, this is not a consequence of the contextualist view, but of the fact that in order to modulate something in a particular context, the "something" that is modulated must exist beforehand.

[61] Observe that the contextual factors *relevant concepts/prototypes* and *weighting of concepts/prototypes*, in the case of the prototype theory, correspond with the factors *relevant exemplars* and *weighting of exemplars* –respectively– in the case of the exemplar theory.

case of the exemplar view– are the locations of the exemplars of the categories that are used in categorizations, inferences, etc. Secondly, in the case of the *classical theory* the modulated thing would be the set of stored rules associated to a certain category, which might be applied or not depending on the context[62]. That would result in a partition of the universe of things between those belonging and not belonging to the considered category, a partition that could vary from occasion to occasion.

In all those cases, the element that undergoes modulation (i.e., prototype, exemplar, or set of rules) may be identified with my notion of *stored concept*, given that they are the only information that needs to be persistently registered by the cognitive system. And, although they are part of the input of the cognitive processes by means of which concepts are instantiated, they do not wholly determine the result of those instantiations in a particular context.

In consequence, even though the discussion in this chapter was focused on the case of the prototype theory of concepts, nothing prevents the same considerations to be applied in approaches based on other theories on the structure of concepts, as long as a contextualist view of concepts is assumed. Therefore, my conclusion is that the mere acceptance of contextualism leads to the need to differentiate two distinct senses of concept, the first associated to the notion of *storage*, and the second associated to the notion of *instantiation*. Additionally, the same can be said of my argument and discussion in section 5.6, that is, since none in them depends on the assumed structure of concepts, the conclusion that instantiated concepts are non-persistent and non-representational is valid for every contextualist approach.

5.8. Conclusions

In this chapter I have firstly tried to show that Casasanto and Lupyan's *ad hoc* cognition framework can be characterized by means of a prototype theory of concepts that deploys a geometrical similarity space. Such a proposal is compatible with Casasanto and Lupyan's thesis that there are no context-independent concepts, and –in the pages above– I have identified four possible sources of contextual dependence (i.e., relevant concepts, kind of metric, importance of dimensions and weighting of concepts).

At that point, and based on the different roles played by regions and prototypes in the conceptual space theory, I also argued for the need to shift the focus from regions to prototypes. That was a significant change, since a definition of concepts in terms of regions –as Gärdenfors does– is misleading, because it could make us falsely attribute to concepts a much more static character than that they really have in the theory. Indeed, if

[62] Clearly, it might be replied that the original definition –call it "definition of level 1"– together with the rules of modulation-by-context constitutes another definition –call it "definition of level 2"–; and that, by virtue of it, we should not speak of modulation, since we could always resort to the (unmodulated) definition of level 2. Anyhow, not everybody would accept that such definition of level 2 could be called "definition". In fact, its acceptance opens the door to the claim that non-classical theories of concepts (e.g., the prototype and the exemplar views) may be reduced to the classical view, to the extent that their associated categorization processes can be expressed by means of a complex set of rules. And the latter is something that very few would be willing to accept.

concepts were identified with static regions within a similarity space, then concepts could not be differently instantiated for each particular context, so –at least– a moderate invariantist position would have to be adopted. By contrast, when the focus is shifted toward the prototypes, then a same concept can be differently instantiated from the same location of the prototype, depending on the rest of contextual factors. (Or, expressed in terms of regions, different conceptual regions may be produced in each particular instantiation, depending on which the relevant context is.)

Next, two different senses of concept were distinguished, associated with two distinct facets in their life cycle. On the one hand, *concepts as storage* were identified with the information persistently registered by the mind about categories, which remains stable between different executions of the concept-acquisition processes. On the other hand, *concepts as instantiation* were said to be the ones responsible of the external manifestation of concepts (i.e., in categorizations and inferences), and were described as non-persistent mental events that happen at the end of those cognitive processes. I also claimed that those two notions of concept should be seen as different facets in the life cycle of a concept, and that this life cycle has a circular structure (what explained the mutual dependence between the notions of *stored concept* and *instantiated concept* for each given category).

One major advantage of this approach is that it brings together virtues both from the contextualist and the invariantist views. In regard to contextualism, my proposal satisfactorily articulates a framework –that of the *ad hoc* cognition– compatible with the evidence against the existence of definitions –or conceptual cores–, in which concepts are construals produced on the fly from a set of occasion-specific cues. The main benefit of this is the framework's capacity to explain our adaptive abilities to changing environments. Regarding invariantism, and putting aside the question of whether it is possible the mutual comprehension of the messages interchanged by the participants in a conversation (an issue that goes beyond the aim of this doctoral thesis), my proposal holds that –in spite of the context-dependence of instantiated concepts– stored concepts are stable enough to accumulate new information on categories.

After arguing for the need to distinguish those two different notions of concept, I held that instantiated concepts should be viewed as non-persistent mental events that result from cognitive processes, not in the sense of *products*, but in the sense of *phenomena* (i.e., non-persistent results that cannot be accessed beyond their occurrence time). This was in line with my thesis that instantiated concepts must not be identified with psychological states –or entities– stored within mental structures, but with something that happens when their instantiation processes end[63]. Next I concluded that, since representations are commonly conceived as relatively stable and persistent objects which encode information, then the instantiated concepts posed by contextualism are not representations, because they do not codify stable information (so they do not fulfill the minimal persistence requirement demanded of any representation).

[63] Or, in other words, all that was in line with the view that *instantiated concepts are non-persistent* mental events that occur in the mind.

Finally, I have shown that, since none of my assumptions in this chapter depends on the chosen theory on the structure of concepts, my conclusions regarding the need of distinguishing between stored concepts and instantiated concepts, and in regard to the non-persistent and non-representational character of instantiated concepts, would apply to any approach that assumes a contextualist view of concepts.

Chapter 6: Empiricist concept acquisition and the circularity threat

Οὗ δ' ἂν τὰ ὅμοια μετέχοντα ὅμοια ᾖ, οὐκ ἐκεῖνο
ἔσται αὐτὸ τὸ εἶδος; Παντάπασι μὲν οὖν.
— *And that by participating in which the likes are like:*
Will not that be the form itself? — *By all means.* –
Plato (*Parmenides* 132e)

6.1. Introduction

One of the main problems of concept empiricism is to account for the acquisition of primitive concepts (i.e., the most basic constituents of concepts) without resorting to preexisting innate elements. According to nativism, an innate representational repertoire (on which later learning is based) must be one of the key elements in any theory of concept acquisition. However, this is not a choice for a consistent empiricist, since it compels to the acceptance that concept learning consists in the production of complex concepts from a set of innate elements. Therefore, a response from empiricism is required to the reasons provided by nativists (Fodor 1975, 2008; Carey 2009) against the thesis that primitive concepts may be learned.

My aim in this chapter is to prove that the uppermost nativist arguments against the acquisition of primitive concepts rest on the assumption that the constituents of concepts must be available beforehand, as an input of the learning process. I call it the *precedence assumption*. Nevertheless, I will claim that there is no obligation to accept such a hypothesis, because the constitutive elements of a concept *C* may result from the same learning process by virtue of which that concept is acquired. To support my perspective, a model of this view is provided. That model will consist in a three-step iterative acquisition system, constituted by two general-purpose abilities (i.e., dimensionality reduction and pattern recognition) and one last stage of evaluation and readjustment of the model.

Having said this, in section 6.2 I introduce the acquisition problem in concept empiricism. There, I firstly place the empiricism-nativism distinction in the present debate, which allows empiricist approaches to incorporate some degree of nativism. In line with this, my (empiricist) proposed model will be compatible with the existence of inherited elements –in the form of a set of (innate) initial seeds– which, in any case, should not be identified with fully developed concepts. After that I remind some of the major difficulties of concept empiricism, and I focus on Fodor's argument –and Carey's reformulation of it– for the thesis that primitive concepts cannot be learned. Next I show that the nativist argument can be generalized, so that it is expressed independently of the kind of learning mechanism –or, more particularly, without assuming that learning is always a process of hypothesis formation and testing–. My conclusion is that there is a circularity threat whenever concepts and their constitutive elements are thought to be acquired through the same kind of cognitive process, because that process ends up with an infinite regress when explaining how the most basic constituents of concepts are learned.

Then, in section 6.3, I hold that the crucial problem of empiricism is not due to the Fodor's or Carey's premise that all learning mechanisms may be reduced to hypothesis formation and confirmation, but to an assumption present in most of empiricist and nativist learning approaches. I will give it the name of the *precedence assumption*, according to which the perceptual or conceptual constituents of a concept *C* have to be available as an input of the process that leads to the acquisition of that concept. In this case my point will be that the acceptance of the precedence assumption constitutes a mistake for the empiricist, since no model built on it can account for the acquisition of general concepts in a non-circular way.

Lastly, in section 6.4 I claim that all is not lost for the empiricist, since it is possible that our repertoire of primitive concepts have been acquired in a non-circular way. The idea is that if the most basic elements of concepts are ready at the end of the learning process –although not from the start–, then the precedence assumption becomes unnecessary and, consequently, the circularity threat –when explaining concept acquisition– disappears. There I will describe an acquisition model built on the basis of a three-step iterative learning system, which is able to produce concepts and their constitutive elements as result of the same execution of the acquisition process. That system will be based on two domain-general computational competences (i.e., *dimensionality reduction* and *pattern identification*), followed by a final stage where the results are evaluated and the system/model is readjusted accordingly. The dimensional reduction module will produce new relevant and reduced factors / dimensions / properties which rule out as much redundant information as possible, while retaining as much variability (of the original data) as doable. The pattern recognition module will search for regularities in the reduced data, and will generate –for an approach based on a cluster analysis– the centroids that, in chapter 5, were identified with the notion of *stored concept*, and from which *instantiated concepts* are produced. Finally, an iterative system is necessary because nothing guarantees that the obtained factors and patterns are the most predictive ones, and this is also the main reason for the (third) stage of evaluation and readjustment of the model.

6.2. The acquisition problem in concept empiricism

Where do concepts come from? As said in chapter 1, this is one of the critical issues facing, not only present cognitive science, but also ancient, modern and contemporary philosophy. Independently of the ontological status attributed to concepts, two major responses to the question on the origin of concepts can be distinguished: (a) *nativism*, according to which many/most/all concepts are innate; (b) *empiricism*, which claims that concepts are a product of experience, and that few of them are innate.

In this section, and after situating the empiricism-nativism distinction in the contemporary debate, which allows empiricist models to accept some degree of nativism, I will show that my (empiricist) proposal is compatible with the existence of inherited elements that, anyway, cannot be identified with fully developed concepts. Then, I recall some of the main problems of concept empiricist, paying special attention to Fodor's argument (and Carey's reformulation of it) in favor of the thesis that primitive concepts cannot be learned. Lastly, I prove that Fodor's argument may be re-expressed in such a way that it is independent of the assumed kind of learning.

6.2.1 The empiricist-nativist debate revisited

One prominent position in the (contemporary) empiricist-nativist debate is the one defended by Margolis and Laurence (2013), who argue that the discussion should not be focused on the innateness of concepts, but on the general-purpose or domain-specific character of the cognitive systems that guide the acquisition of those concepts[1].

In Margolis and Laurence's view, both nativists and empiricists are committed to innatist assumptions. On the one hand, the main thesis of nativism is that most, if not all, concepts are innate. On the other hand, it is also true that empiricists tend to accept the existence of innate processes, mechanisms or dispositions, on the basis of which concepts are acquired. Therefore, it seems that both nativists and empiricists include innate elements in their cognitive models.

Nevertheless, that is only one of the two hypotheses that characterize a position which can be wholly described as follows:

[A] Both nativists and empiricists accept the existence of innate learning systems so, in consequence, nativism cannot be defined by the thesis that most of concepts are innate (in the sense of being acquired by means of an innate psychological base).

[B] Both nativists and empiricists may be disposed to accept that (some) concepts can be learned from experience together with a certain acquisition base. Therefore, empiricism cannot be defined by the thesis that concepts come from the accumulation of sense experience.

Their point is that, if the nativism-empiricism distinction cannot be based on the innateness of the underlying acquisition systems, nor on the acceptance that concepts are learned from experience, then it may be thought that the focus should be put on the different character of the cognitive systems which constitute the acquisition base.

On this basis, Margolis and Laurence hold that the main point of disagreement between nativism and empiricism is not about the existence or not of innate cognitive elements –because both nativists and empiricists accept them alike–, but about the specific or general character, respectively, of such innate systems. Thence, according to them it is necessary to distinguish the nativist acquisition base, which would be constituted by a set

[1] This is a fairly widespread view in the contemporary debate (Godfrey-Smith 1996; Spelke 1998; Cowie 1999), which can be traced back to early modern rationalism (Descartes 1647; Leibniz 1765).

of domain-specific modules, from the empiricist acquisition base, which would consist in a set of general-purpose systems.

From here on, I will assume Margolis and Laurence's distinction between nativism and empiricism, based on the character –specific or general, respectively– of the underlying acquisition systems. In particular, my proposal in this chapter fits in their definition of empiricist approach, because it is based on a set of general-purpose acquisition systems. At the same time, it does not conform to other stronger views of empiricism, since I accept both that those learning systems are innately given to the subjects, and that some innate informational elements can intervene in them –as I will describe in the following section–.

6.2.2 Innate elements in empiricist models

According to Carey (2009; 2011), any theory of concept development should address the three following questions: (a) the innate conceptual repertoire; (b) the differences between such innate concepts and the ones of an adult; and (c) which learning processes lead to the transformation of the first into the latter.

And, even though my (empiricist) proposal is very different from Carey's nativism, on my view she is right when claiming that the three mentioned issues have to be addressed by any theory on the acquisition of concepts. Hence, my work in this chapter deals with the problem of concept learning from an empiricist perspective, taking care of some innatist demands. The idea will be to accept innate elements in the empiricist model in such a way that they do not constrict the model, nor prevent its adaptation to changing environments.

In regard to the first issue –i.e., the innate conceptual repertoire–, and though my model will fit in the empiricist cannon, according to which the only innate structures are general purpose mechanisms on whose basis the acquisition of any concept may be explained (Laurence and Margolis 2002, p. 51)–, my proposal can also include an innate repertoire in the form of a set of (innate) initial seeds used by the pattern recognition module –which will be characterized by means of cluster analysis–. Anyhow, such initial seeds[2] are not genuine concepts, but *pre-concepts* tentatively used as a starting point by the cognitive processes that lead to the identification of patterns then associated to concepts. Thus, although the role played by those initial seeds is far from the common notion of (innate) concept –because they are just a vector to guide the pattern recognition processes–, they might explain why different people produce similar concepts for many categories. Additionally, since the initial seeds do not have to be associated to a particular set of dimensions[3] (for instance, in a conceptual space), I am not accepting a particular innate conceptual repertoire which could hardly account for conceptual variability from one person to another and, also, from one generation to the next.

[2] Or locations of the first centroids for the processes of cluster analysis.

[3] Indeed, my thesis is that –in general– we inherit the initial seeds that guide the acquisition process, but not the dimensional space within which those seeds are located. Notwithstanding this, I agree that in some cases dimensions can also be inherited when they are very close to sensory experience, as happens in the case of colors, shapes, etc.

Obviously, the idea of (innate) initial seeds in an optimization process differs from the notion of innate concept which some empiricists could –at best– accept (that is, it is very different from an innate stock of conceptual primitives restricted to sensory or perceptual representations) (Piaget 1937; Quine 1969). And, it is also distinct from the innate conceptual representations accepted by moderate nativists, even when the latter are conceived as the product of (innate) input analyzers which are not the result of learning (Carey 2011, 2015). By contrast, the kind of innate element considered in my model is a much weaker notion, since it is not directly related to a particular set of dimensions, not even through an innate input analyzer. Consequently, my view is neutral on the issue of whether learning in a given domain is based on innate information about that domain. Therefore, although an approach like mine is compatible with the existence of innate elements which guide the learning of concepts, it does not require them in order to provide an explanation of how concepts are acquired (so, in this respect, my proposal is maximally flexible). Indeed, my perspective is consistent with the empiricist hypothesis that subjects do not inherit fully developed concepts, and also with the nativist evidence that there is a kind of biological preparedness to quickly acquire cognitive abilities for carrying out mental tasks crucial from an evolutionary point of view (Seligman 1971; Cummins and Cummins 1999).

As a result, and in regard to Carey's second issue, the differences –in my view– between the innate elements and the concepts of an adult are significant. Firstly, I reject that the aforementioned (innate) initial seeds are genuine concepts, because –as said above– they are not framed in a certain set of dimensions. Secondly, those seeds need to be both specifically adapted to the subject's experiences, and enclosed in a dimensional framework that allows applying them in cognitive tasks (i.e., in categorizations). Finally, in connection with Carey's third question, section 6.4 is devoted to the description of the acquisition processes which lead to the production of final concepts from experience information and a set of innate (and tentative) initial seeds.

6.2.3 Empiricism and its opponents

As stated in chapter 1, perhaps the most serious objection to concept empiricism are Fodor's (1975, 1981) arguments for nativism –and against empiricism–. According to them, concepts cannot be learned since their constituents should be available beforehand the acquisition process (in order to be used as input of the computational apparatus), what ends up with an infinite regress when trying to explain how the constitutive features of concepts are acquired[4]. And, although I deal with Fodor's argument in section 6.2.4 below, now it is worth considering how that criticism has been applied to contemporary empiricism.

[4] Another distinct, although related, problem is that of how the constraints that guide the learning process are acquired. In regard to this, it may be thought that it is the same problem as the one that rose from Fodor's arguments against empiricism, to the extent that it seems necessary a non-circular explanation of how those learning restrictions are initially acquired. Having said this, in this chapter I will focus on Fodor's arguments, since I consider that the sort of response provided here to these latter can also be applied to the case of the learning constraints.

Let me consider, for example, Marcus' (1998) critique against the proposal of Elman *et al.* (1996) that connectionist models can explain the acquisition of new concepts without resorting to a previous innate repertoire[5]. On my view, the most interesting aspect of Marcus' critique is that, even though his work is designed against the connectionist approach, he focuses his arguments on the most problematic element of concept constructivism (and, consequently, of concept empiricism[6]).

Marcus' main idea is that the largest problem of constructivism is that there has never been a concrete computational explanation of how a learning system could produce constructivist conceptual redescriptions –that is, new concepts not based on a previous set of already existing concepts–. His conclusion is that Elman *et al.* do not prove that connectionism is able to provide that kind of explanation, because the input and output schemes used by their connectionist models are also concepts. Indeed, the underlying problem identified by Marcus is the potential inability of the empiricist models to produce real conceptual emergence (i.e., their potential failure to "develop new concepts where there were none") (Marcus 1998, p. 161). The key issue is that most constructivist models do not produce new concepts, but correspondences between predefined sets of concepts, which are presupposed by all those models[7]. This is ultimately the issue that underlies Fodor's main argument against empiricism.

Certainly, Marcus' point is closely connected with the main argument for nativism, and against empiricism, already mentioned in chapter 1, according to which the stimuli from environment are so poor that concepts could not be produced merely from them. Indeed, the poverty of the stimulus argument ultimately is an objection based on the lack of a concrete cognitive model, powerful enough to process stimuli richer than commonly expected[8]. Here the point is that, if such a model existed, then the domain-specific systems proposed by the nativist could have been acquired by means of domain-general mechanisms, and that would tip the balance to the empiricist side.

Nevertheless, when the empiricist tries to formulate concrete cognitive models to explain the acquisition of new concepts, he stumbles on the constructivist problem of producing new concepts without assuming a set of predefined and primitive elements. As Marcus argued, the main constructivist/empiricist problem is that they do not explain how a new concept can be acquired from none. And, again, this is essentially Fodor's argument against concept empiricism.

My purpose in this chapter is to show that a multivariate-analysis learning system, based on a combination of (a) *dimensionality reduction* –conceived in terms of a factor analysis or principal component analysis (PCA)–, (b) *pattern identification* –envisaged as

[5] For a previous and clearer discussion –from connectionism– against Fodor's nativist argument, see Molenaar (1986).

[6] Throughout the rest of this section, everything said about constructivism will also apply to empiricism.

[7] And although Marcus recognizes that some constructivist models (Hinton *et al.* 1995) seem to be able to explain the unsupervised acquisition of new concepts (in the form of output units in a neural network), he then dismisses them because of not being robust enough to explain the acquisition of general features across a wide range of environments.

[8] For instance, this was precisely Laurence and Margolis' (2001) critique to Cowie (1999).

a cluster analysis–, (c) and a set of end conditions, is a kind of domain-general cognitive architecture able to produce new concepts in an unsupervised way, without presuming the existence of a previous set of innate conceptual elements. As said above, this approach is compatible with the existence of innate, although not dimensionally structured, information (e.g., the seeds of a cluster analysis, which might be inherited and stored as mere series of numbers[9]). Nevertheless, before dealing with these issues, I will carefully analyze Fodor's argument –and Carey's reformulation of it–, and try to show that one common hypothesis underlies to any generalization of that kind of argument.

6.2.4 Fodor's and Carey's arguments

According to nativism, an innate representational repertoire –on which later learning is based– is one of key elements in every theory of concept formation. In other words, concept learning entails the production of complex concepts based on a set of other primitive and innate conceptual elements. This view is sometimes called the *building blocks* model of concept learning and, as stated by Margolis and Laurence (2011b), it is a perspective openly accepted both by radical nativists (Fodor 1981), and by moderate nativists (Pinker 2007); and also tacitly assumed by almost all empiricist explanations of concept acquisition.

Therefore, by virtue of the definition of concept empiricism –according to which all concepts are learned, not innate–, every empiricist theory has to face the problem of providing an account of the acquisition of general concepts, without resorting to a preexisting innate repertoire. (In other case, it would be some kind of moderate nativism, but not an empiricist theory of concepts.) Or, in other words, it is required an answer to the reasons provided from nativism against the thesis that primitive concepts can be learned. As it is well known, Fodor has cogently argued against the acquisition of primitive concepts, and –less convincingly– against the learning of complex concepts[10]. The aim of the present section is to review the argument provided by Fodor against the acquisition of concepts, and also Carey's reformulation of it.

Fodor's argument against empiricism

According to Fodor, every empiricist theory of concept learning is tacitly committed to some kind of nativism. However, Fodor's position has not remained unchanged, so different versions of his argument against concept empiricism may be found throughout his work[11].

[9] Consequently, these seeds do not need to be associated to any particular dimensional space.

[10] The other main conclusion of Fodor's works against empiricism is that most concepts do not have any internal structure. This would explain the lack of success of the research program carried out by empiricists in order to reduce complex concepts to their primitive constituents. On my view, Fodor's interpretation of the failure of the reductionist program is not mandatory, since such a failure could also be due to the influence of context on any application –or instantiation, in my terminology– of a concept.

[11] Critical discussions of Fodor's arguments –and of their consequences– from other points of view may be found in Molenaar (1986), Samet and Flanagan (1989), Boom (1991), Margolis (1998), Laurence and Margolis (2002), Recanati (2002), Margolis and Laurence (2011c), and Rey (2014).

The earliest version of this argument goes back to *The Language of Thought*, where Fodor argued for the impossibility of changes in the conceptual system of an organism (i.e., in favor of the thesis that a conceptual system cannot be produced by a weaker one by means of hypothesis formation and confirmation). Such an argument could be summed up as follows (Fodor 1975, p. 93; 1980, p. 148):

[P_1] Learning is a process of hypothesis formation and confirmation (HF).
[P_2] In order to formulate a hypothesis on a concept, the learning system must be able to represent such a concept.
[P_3] A concept cannot be represented by a system before that concept is learned.
[P_4] Thence, concepts –either primitive or complex– cannot be acquired by HF.
[P_5] There is no other general learning method (Fodor 1975, p. 58; 1981, p. 269).
[C_1] Concepts cannot be learned.

Before continuing, one clarification is needed regarding proposition [P_4] in this first argument. In latter works Fodor is more conservative, and restricts proposition [P_4] to the case of primitive concepts[12] (Fodor 1981, pp. 270-272). In such a work, Fodor expressed that proposition in the following sense:

[P_4'] Primitive concepts cannot be learned by HF. The primitive concept C_p being learned is the concept whose acquisition has to be explained, so C_p will have to be available in order to formulate the tested hypothesis, even in order to identify the experiences fixed by such a concept. Therefore, the hypothesis required for the learning of C_p cannot even be formulated.

Hence, attending to Fodor's previous cautions, proposition [P_4] may be thought to be not applicable to complex concepts since, in the case of the acquisition of a complex concept C_c, the hypothesis might be formulated in terms of an empty concept –which would act as a void label (e.g., *stimulus-1*)–, and of the rest of primitive concepts constituting the observational statement (e.g., *stimulus-1 is green*). This empiricist reply is not possible if the learned concept is primitive, so Carey's subsequent restriction of [P_{III}] to the case of primitive concepts seems to be a good decision[13].

All this considered, a second weaker version of the argument –that is, the one for primitive concepts– may be expressed as follows:

[P_1'] Learning is a process of hypothesis formation and confirmation.
[P_2'] If [P_1'], then to learn a concept C is to produce a hypothesis "x is a C iff x is $f_1, f_2, \dots f_n$" (HF-rule), and to check it against experience.
[P_3'] In order to begin the learning process, the system has to be able to represent / recognize the constitutive elements $f_1, f_2, \dots f_n$ of C.
[P_4'] Concepts $f_1, f_2, \dots f_n$ (or concepts in the right side) can be learned or unlearned.

[12] Even though later in his work –consider, for instance, *LOT2* (Fodor 2008)–, Fodor re-extended again proposition [P_4] to complex concepts (as he did in this first version of the argument).

[13] See her argument below in this section.

[P_5'] Not all concepts available for the right side may be learned, since HF requires projecting a HF-rule. (Otherwise, concept learning could never begin.)
[C_1'] Concept learning presupposes an innate stock of concepts.

Thus, Fodor concludes that the empiricist thesis that concepts result from assembling their definitional constituents requires the existence of an innate repertoire of primitive concepts, which contradicts the empiricist hypothesis that most, if not all, concepts are learned. In this case, Fodor's conclusion is not that concepts –both primitive and complex– cannot be learned (that was the conclusion [C_1] of the first version of his argument); but only that primitive concepts cannot be learned, since they are presupposed by the empiricist acquisition models.

Finally, a third version results when the second argument is extended from the conclusion [C_1'] –or premise [P_6'] in the third argument– that primitive concepts cannot be acquired, to the conclusion that complex concepts are also innately specified. Such an argument can be expressed as follows:

[P_6'] Primitive concepts cannot be learned, and thus must be innate.
[P_7'] The set of all available concepts is the closure of the primitive ones under the combinatorial mechanisms[14] (which are supposed to be innate) (Fodor 1981, p. 264).
[C_2'] So, every potential complex concept is innately specified.

According to this third version of the argument, any complex concept is a function of a set of innate primitives, and of what can be built from them by means of the logical resources available to the learner[15]. Therefore, the expressive power of our conceptual system would be innately determined, and could not be increased through learning.

Carey's reformulation of Fodor's argument

Some years later, Susan Carey re-expressed Fodor's argument for the thesis that primitive concepts are innate (or, in other words, for the thesis that primitive concepts cannot be learned) in a much clearer way[16] (Carey 2009, p. 513):

[P_{I}] All learning mechanisms reduce to hypothesis formation and testing.
[P_{II}] Hypotheses that play a role in learning new concepts need to be formulated in terms of the available concepts, and the principle of compositionality.
[P_{III}] Primitive concepts are not formulatable definitionally –nor probabilistically– in terms of other concepts.
[C_{I}] Therefore, primitive concepts cannot be learned.

[14] Or, in other words, the expressive power of our conceptual system is determined by its primitive concepts and the combinatorial principles that rule them.

[15] A similar kind of discourse is present in many others, like Rips *et al.* (2008) and Rey (2014).

[16] From here on, I shall mainly focus on Carey's reformulation of Fodor's argument.

At this point, it is worthwhile to recall Carey's own view regarding her formulation of Fodor's argument[17]:

> Learning is a computational process that operates on representations, and therefore if we believe that learning plays a role in the construction of representational resources, we are committed to there being *some* innate representations: those that are the input to the initial episodes of learning. (…) Remember, I am using "innate" to mean "not learned." (Carey 2011, p. 152)

Carey's idea is that, because learning operates on concepts, and given that concepts are the result of learning processes, then there must exist some primitive concepts used as an input of, at least, the first episodes of learning. And, since those primitive concepts cannot be produced by the referred acquisition processes –due to circularity reasons–, then it can be said that they are innate (or, in other words, that they are not learned). This is an important issue, seeing that Carey explicitly states an assumption which, as I hold in section 6.3 (where I call it the *precedence assumption*), is implicitly present in all models of concept learning.

In regard to Carey's formulation of the argument, premise [P_I] is equivalent to the conjunction of the premises [P_1] and [P_5] in Fodor's first version of the argument[18]. In fact, premise [P_I] –or a variation of it– plays a key role in any of Fodor's and Carey's arguments. And, although the premise that all learning mechanisms may be reduced to hypothesis formation and testing (i.e., [P_I] and its variations) has been severely criticized –as I describe in the next subsection–, my thesis will be that HF is not a condition required in order to build this kind of argument against concept empiricism (see section 6.2.5 below).

HYPOTHESIS FORMATION AND TESTING

As shown in the previous pages, although Fodor's arguments against empiricism –and the thesis that concepts are the result of assembling their most basic elements– have slightly changed over time, all of them can be fairly summarized by the next scheme:

[P_i] Concepts, either primitive or complex, cannot be acquired by hypothesis formation and confirmation.
[P_{ii}] There is no other learning method.
[C_i] Thence, concepts cannot be learned.

Regarding premise [P_{ii}] –which is equivalent to Carey's premise [P_I]– Fodor has traditionally argued for it as follows: every theory of concept learning requires the inductive fixation of beliefs, a task in which the formulation and confirmation of hypotheses is a must (Fodor 1975; 2008). The present section is devoted to this assumption.

[17] The reader must keep in mind that there where Carey says "representation" I will read "concept", even though –as said in chapter 5– in my view (instantiated) concepts have non-representational character.

[18] In subsequent versions of the argument, Fodor did not make premise [P_5] explicit, so it can be said that premise [P_I] in Carey's formulation is also implicitly equivalent to premise [P_1'] in Fodor's second and third versions of it.

One of Fodor's main reasons in favor of HF is the claim that learning mechanisms have to be rational, which allows to support a kind of internalist justification, against other authors –like Margolis and Laurence (2011c)– who think that an externalist justification is enough. Leaving aside the issue of justification, Fodor holds that HF is the only type of learning where concepts are rationally acquired. However, on my view the focus on rationality is misguided. Most authors would subscribe that learning is a computational process, but computation is neither rational nor non-rational. Indeed, computation is mere calculation. *Fast*, *efficient*, *optimized*, and others are adjectives that could apply to computation, but *rational* is not one of them, and the same can be said about *non-rational*. Therefore, the rationality –or non-rationality– of a cognitive process is not a reason to accept –or discard– it as a potential learning mechanism[19].

Additionally, Fodor's arguments have been criticized by moderate nativists, such as Margolis and Laurence (2001b), who claim that both premises, [P_i] and [P_{ii}], are mistaken. Firstly, premise [P_i] used to be questioned for the case of complex concepts. But, this is in line with the doubts brought up by Fodor himself on that issue, which moved him back and forth from one version of the argument to another. Because of this, and since –as said above– I will focus my discussion on Carey's reformulation of Fodor's view (which is constrained to the case of primitive concepts), no more will be said on these objections to premise [P_i].

Secondly, premise [P_{ii}] has been put into question by many who think that not all learning mechanisms may be reduced to hypothesis formation and confirmation. For instance, Margolis and Laurence (2011c) argue that [P_{ii}] is patently wrong for the case of complex concepts since, for them, there are many other kinds of learning different from HF, such as rote learning, communication-based learning, or automatic associative learning. In regard to this critique, it should be noticed that it is addressed against the argument for complex concepts, so they have a minor influence in the argument constrained to primitive concepts.

Even worse, it is not clear whether any of those alternative learning methods both accounts for the general acquisition of concepts –which is the issue disapproved by Fodor–, and is free from HF:

- In respect of *rote learning* and *communication-based learning*, it is doubtful that these methods lead to the acquisition of concepts in a general sense, because in both cases the learned concept is fixed by a list of instructions (or a set of features) provided by other subject. In this case Fodor could reply that this is not the spontaneous (unsupervised and sub-personal) type of learning that seems to lead to the acquisition of most of concepts.
- As regards *automatic associative learning* (AAL), it is the only alternative method versatile enough to explain the acquisition of concepts (not objects), in such a way that their constitutive properties are not to be explicitly given by an instructor. In this case the problem is that, there is evidence that AAL could be a case of hidden HF. Indeed, according to Yu *et al.* (2007), both HF and AAL could be special cas-

[19] For a discussion about whether "rational" is, or not, a notion that may be applied to cognitive subpersonal processes, see Evans and Stanovich (2013, p. 229).

es of the very same set of learning mechanisms, since the election of the strongest associations in AAL may be identified with the formulation of hypotheses in HF.

Having said all this, on my view there are still strong reasons for Fodor's premise [P_{ii}] –or Carey's premise [P_{I}]–, at least in the case of primitive concepts. Notwithstanding this, its acceptance is not necessary to make empiricism get into trouble since, as will be shown in section 6.2.5, Fodor's and Carey's arguments can be generalized so that they are independent of the kind of learning.

6.2.5 Generalization of the criticism

As noted in the last section, the nativist argument against the thesis that concepts can be learned may be formulated in multiple ways. Nevertheless, all those distinct expressions depended on the controversial thesis that all learning can be reduced to HF.

My aim here is to show that Carey's formulation of Fodor's argument can be generalized, so that it is expressed independently of the assumed kind of learning mechanism –or, in other words, without presuming that there is no other learning methods different from HF–. Such a generalization may be expressed as follows:

[P_{I}'] Concept acquisition is the result of learning processes.

[P_{II}'] The perceptual or conceptual constituents of concepts should be available before the beginning of the learning process, in order to be used as an input of its computational apparatus.

[P_{III}'] Thence, those constituents should have been acquired at an earlier time.

[P_{IV}'] A model **M** like this cannot explain the acquisition of such constituents without falling into circularity[20]. (Those constituents should result from a learning process in whose beginning another set of more basic elements must be available, and so on and so forth.)

[C_{I}'] Therefore, primitive concepts cannot be learned.

The structure of this generalization is analogous to the one of Carey's argument. First, complex concepts result from a certain kind of learning (i.e., main empiricist assumption), so premise [P_{I}'] generalizes premise [P_{I}]. Second, their constituents should be available before the beginning of the learning process, so [P_{II}'] generalizes premise [P_{II}]. But, premises [P_{I}'] and [P_{II}'] lead directly to the circularity problem, through propositions [P_{III}'] and [P_{IV}'], which were summarized by Carey in one only proposition [P_{III}]. (The issue is that there is a circularity threat if both concepts and their constitutive properties are acquired by means of the very same kind of cognitive process, since the process ends up with an infinite regress when trying to explain how the constitutive elements of concepts are acquired.) Lastly, both conclusions are exactly the same.

And because premises [P_{I}'] and [P_{II}'] are weaker than premises [P_{I}] and [P_{II}], respectively, this later formulation of the argument against empiricism is more general than those of Fodor and Carey, and, consequently, includes all the cases included by them.

[20] And the same can be said about any other model **M'** which included **M** (and, consequently, were bigger than **M**) and functioned under the aforementioned conditions.

6.3. The precedence assumption

Now, in this section I will prove that the crucial problem of empiricism is not due to Fodor's premise [P_1] –or Carey's premise [P_I]–, that all learning mechanisms can be reduced HF (which led to the innateness of both primitive and complex concepts), but to an assumption common to most of empiricist and nativist learning models.

Indeed, once the nativist argument is generalized, as I did in section 6.2.5, it is clear that the origin of the circularity problem is the premise [P_{II}'] or, more specifically, the *precedence assumption* underlying [P_{II}']:

[PA] The perceptual or conceptual constituents of concepts must be available as an input of the cognitive processes that lead to the acquisition of those concepts.

This is a common assumption of both empiricist and moderate nativist approaches. Certainly, that is the very precise idea which underlies Fodor's reasons against concept empiricism, when he argues that if a primitive concept C_p is available to a subject *S* for hypothesis formation, then *S* already has that concept, since in the absence of C_p the subject would not able neither to formulate the tested hypothesis nor to recognize the experiences[21] associated with C_p.

The reasons provided by many other authors for their tacit acceptance of [PA] are quite similar to those of Fodor. For example, when Carey expresses her view regarding this point, it seems clear that she assumes that a set of representations is required as an input of the learning process –or, in other words, she takes for granted the [PA]–:

> (...) if we believe that learning plays a role in the construction of representational resources, we are committed to there being *some* innate representations: those that are the input to the initial episodes of learning. (Carey 2011, p. 152)

However, this common sense assumption –seemingly trivial and unproblematic– is a mistake for the empiricist[22], since no model built on it can provide an explanation of the acquisition of general concepts free from the circularity threat. The consequence seems to be that under the [PA], and in order to avoid circularity, the empiricist has to accept the innateness of primitive concepts, which explains the usual shift of empiricism towards moderate nativism.

[21] In this second case, I think that Fodor is essentially right because, in order to identify and manage the observational experience(s) linked with C_p, the perceptual data associated with C_p should have been previously stored within our cognitive system.

[22] Nevertheless, the [PA] is a problem not only for the empiricist, but also for the moderate nativist, because once it is concluded that primitive concepts cannot be learned, this may lead, through proposition [P_7'] –see section 6.2.4 above–, to the radical nativist conclusion that every potential concept in our cognitive system is innately determined.

6.4. A non-circular model for concept learning

My aim in this section is to prove that all is not lost for the empiricist, since it is possible that our repertoire of primitive concepts has been acquired in a non-circular way. My thesis will be that it is enough that the most basic elements of concepts are ready at the end of the learning process –although not from its beginning–, which avoids the precedence assumption and, in consequence, the circularity threat when explaining concept acquisition.

On that basis, I propose a learning model where the constitutive elements of a concept can result from the very same cognitive process by virtue of which that concept is acquired. That model will consist in a three-step iterative learning system, constituted by two kind of general-purpose learning abilities, to wit, *dimensionality reduction* and *pattern recognition*; followed by one last stage that evaluates and readjusts the model.

6.4.1 A different kind of learning

At this point, should the empiricist give up and accept that primitive concepts cannot be learned? From my point of view, there is a way out for empiricism, since there is no obligation to accept [PA]. In this section I will claim that a different kind of learning is possible. Thence, the thesis I will argue for is that, for a model (of concept acquisition) to be able to function, it is not necessary that its constitutive elements are available from the very beginning. In fact, my proposal is an empiricist acquisition model where the most basic constituents of a concept result from the very same learning process by virtue of which such a concept is acquired. Consequently, this kind of learning model could explain the acquisition of general concepts (i.e., both of concepts and their primitive constituents) in a non-circular way, that is, without resorting to a preexisting innate repertoire of primitive concepts.

The model consists in a three-step iterative acquisition process, constituted by two kinds of general-purpose learning abilities (i.e., dimensionality reduction and pattern identification[23]), which will be characterized by means of multivariate analysis (MVA) techniques (see Fig. 6.1):

(1) *Dimensional reduction*: since a reduced perceptual input will be more efficiently stored and processed by the subsequent cognitive systems. This first step will produce the most basic constituents of concepts.

(2) *Pattern recognition*: regularities in the (reduced) data input are singled out in order to pinpoint similar future stimulus in comparable circumstances. Such regularities may be identified with the concepts of our mental system.

(3) *Evaluation and readjustment of the model*: the resulting dimensions and patterns –or alternatively, the resulting features and concepts– will be evaluated in terms of their predictive power and, based on it, the iterative learning process will be rearranged accordingly.

[23] These two learning modules are conceived to work in an autonomous and unsupervised way.

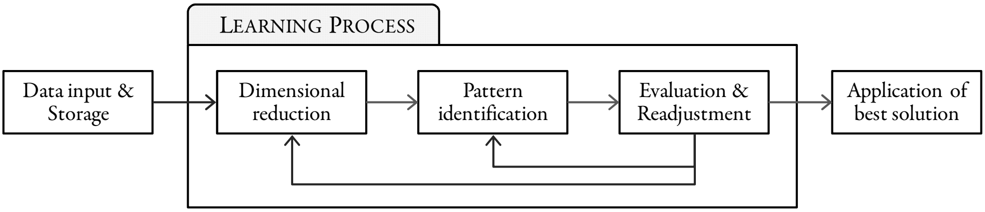

Fig. 6.1. General structure of a non-circular model for the acquisition of complex and primitive concepts. The model is constituted by a three-step iterative learning process formed by three sequential stages: (i) dimensional reduction; (ii) pattern recognition; and (iii) evaluation and readjustment of the model.

This scheme is inspired by the traditional view in pattern recognition literature, of dividing the learning system into a feature extractor (*dimensionality reduction*) followed by a classifier (*pattern identification*) (Jain *et al.* 2000; Webb 2002; Fichet *et al.* 2011).

In the literature on concept acquisition there are works which make use –at least partly– of some of those three elements. On the one hand, and though the application of MVA techniques to concept learning and categorization is common (Diday 2005; Napoli 2005; Vanpaemel and Storms 2010), the general structure of those approaches does not usually include a previous dimensionality reduction –in contrast to the one here proposed– to explain the acquisition of new features. On the other hand, many times the cognitive abilities intervening in concept acquisition are conceived as modules working one after another, and not as different stages within one unique iterative learning process. And regarding those proposals which conceive concept learning as an iterative model (e.g., Bloch-Mullins 2018), they usually –and tacitly– accept a set of preexisting representations and, due to it, they do not account for how such innate elements could have been learned. As a consequence, none of those approaches is able to provide an answer to Fodor's circularity objection[24].

6.4.2 Basic assumptions of the model

In the following sections I provide a detailed description of the motivations, articulation and strengths of the three main elements of my proposal, explaining why this approach constitutes a non-circular explanation of concept learning.

However, before beginning any exposition, I will make explicit the assumptions of this model, paying special attention to those regarding the most basic input provided by our perceptual modules, as well as those associated to the issue of how the perceptual input can be temporarily stored –before moving on to the acquisition system–:

[24] By contrast, neural networks seem to fall under the same group as my perspective, since they are able to extract –in an unsupervised way– statistically relevant features by means of self-organizing learning mechanisms (Von der Malsburg 1973; Kohonen 1982; Ritter and Kohonen 1989; Schyns 1991), so they would not be committed to [PA] and would not be threatened by circularity. Unfortunately, neural networks face to the hard problem of interpreting their results, which is something much less tricky in the case of a MVA-based approach like the one here proposed.

(I) Let it be a set of *n* perceptual inputs, $\{a_1, a_2, \ldots a_n\}$, extended over *n* continuous domains. This point is quite uncontroversial, since sensorium provides analog perceptual outputs over continuous (electrochemical) potential levels.

(II) Additionally, I will assume that all the information provided by the sensorium is stored as raw data. This presumption is unproblematic both in terms of *what* to store –because everything is stored–, and in terms of *how* to record the data –since it is enough to register a continuous value, which may be done over an analog storage medium (e.g., an electric potential level or difference)–.

My point here is that those raw data may be recorded and managed without requiring a previous conceptual structure. However, the data should be stored in a way such that it reflected the simultaneity (or non-simultaneity) and temporal order of storage[25], for example, by means of a kind of time stamp[26].

(III) The model is assumed to be constituted by two kinds of innate general-purpose learning mechanisms, namely, a system able to carry out a reduction in the number of dimensions of the raw perceptual input, and a system able to recognize regularities within data sets. Both these abilities are purely computational competences –that is, they merely operate on numbers–, and hence do not require the existence of a prior conceptual system.

And, although in my work those two abilities (i.e., dimensional reduction and pattern recognition) are articulated through MVA techniques (i.e., factor analysis/PCA and cluster analysis, respectively), they can also be conceived in terms of other alternative approaches. For instance, both the dimensional reduction and the pattern recognition stage may be described by means of neural networks (Carpenter 1989; Bartsch 1996; Hinton and Salakhutdinov 2006)[27].

(IV) Finally, the system in charge of the evaluation and readjustment of this iterative learning process is also assumed to be an innate cognitive module in our mind.

I agree with Fodor, and also with many contemporary empiricists, that most –or even all– of the aforementioned systems may be considered innate; from the perceptual systems (and the subsequent modules carrying out the automatic post-processing of the perceptual input), to the cognitive systems that articulate both dimensional reduction

[25] This temporal information is crucial when looking for temporal associations within the raw data –i.e., diachronic regularities over which evaluate the results of the model–.

[26] Time stamps may be conceived as electric potentials stored by analog devices with an exponential decay rate. A time stamp like this would be very precise –in relative evaluations– for the short term, and much less accurate for registers stored long time ago.

[27] Other approaches to dimensionality reduction would be the non-negative matrix factorization (NMF) (Lee and Seung 1999; 2001), and the singular value decomposition (SVD). For a review of the history of the SVD, see Stewart (1993).

By contrast, multidimensional scaling (MDS) cannot be said to be a real alternative to factor analysis / PCA, as a method for dimensionality reduction. As said in section 3.3.2, MDS allows reducing the dimensionality of information, but it uses as starting point the dissimilarities/distances between the considered objects. Unfortunately, that is not the kind of information that can be thought to available as an input of our acquisition processes.

and pattern recognition, I am disposed to accept that those systems are surely innate[28]. Nonetheless, this has no influence on my view, given that the two intervening learning systems (i.e., those associated to dimensional reduction and pattern identification) are general-purpose mechanisms so, according to the nativism-empiricism distinction proposed by Margolis and Laurence (2013) (see section 6.2.1 above), my perspective qualifies as an empiricist model of concept acquisition.

6.4.3 Dimensionality reduction

Concept learning is commonly identified with the recognition of patterns and similarities in the objects of the world. Nevertheless, many times little is said about how the constitutive properties/dimensions of those objects are determined. That is called the selection problem (Goodman 1972; Machery 2009). And, even worse, the size of the data provided by sensorium is so huge that it is hard to imagine how they can be processed without a previous informational reduction.

The aim of this section is to prove that dimensionality reduction constitutes an answer to both of those questions. On the one hand, it explains how redundant information can be ruled out from perceptual input. On the other, it makes clear how the relevant dimensions might be chosen –in an automatic and unsupervised way–, which solves the selection problem[29].

Reasons for a reduction of dimensions

Many times it is suggested that the domain-general learning systems presumed by empiricism are highly sophisticated statistical mechanisms, able to recognize subtle regularities and patterns (Prinz 2002). All those approaches have to face the issue of how to analyze the information provided by the sensorium. Here the problem is that the amount of perceptual data is so huge that it cannot be persistently stored in memory, nor directly analyzed by the pattern recognition mechanisms. This is so since the task of finding statistical relations becomes more difficult as the number of dimensions increases; and the dimensionality of perceptual data may be very high. A way of addressing this issue is by means of a reduction in the number of dimensions which describes most of the original data variability in terms of a reduced number of factors, removing as much redundant information as possible[30, 31].

[28] In regard to the third stage of my proposal (i.e., that associated with the evaluation and readjustment of the model), I prefer to remain silent about its innate or learned character.

[29] For a detailed description of how dimensionality reduction constitutes an answer to the problem of feature selection and extraction , see Webb (2002, ch. 9).

[30] *Dimensionality reduction* is a special kind of *information reduction* which, as a practice largely established in cognitive psychology (Posner 1964) and neural computation (Redlich 1993) –at times also called *information compression* (Wilkenfeld, forthcoming)–, is carried out by means of a transformation that produces an output informationally smaller than its associated input. (Even though concept learning is sometimes identified with an information reduction (Hunt 1962), it is more accurate to say that some intervening modules in concept acquisition play an informational-reduction role.)

The idea is that, given that many sensorial inputs are highly correlated, it is feasible a description of most input data variability by means of fewer dimensions. In other words, if an aggregation of external stimuli *e* (susceptible to be identified with an object or a concept) is composed by *n* perceptual inputs, such a combination of stimuli can be characterized –and, subsequently recognized– through a space with dimensionality much smaller than *n*. This view is a common idea in the analysis of real images, where each image is decomposed in a set of constitutive elements, which are in turn constituted by particular sets of features[32] (Morgan 2016). Inasmuch as the final number of variables is fewer than the number of variables in the original data, its processing will be less resource demanding in terms of memory and computational power. Additionally, a dimensional reduction provides a set of relevant dimensions (free from redundant information), which will increase the effectiveness of the classifiers and analyzing processes working on them.

Obviously, if dimensional reduction happened at the beginning of the acquisition processes, all the rest of the considered cognitive systems would work with the transformed –and reduced– input, not with the original one.

Factor analysis / Principal component analysis

Factor analysis (FA) and principal component analysis (PCA) are two different MVA techniques used to determine the correlation structure among a set of *n* observed variables x_1, x_2, ... x_n, in order to describe their variability in terms of a lower number *k* of unobserved variables –called factors or components, respectively–. Thence, both FA and PCA can produce a reduction in the number of dimensions of the input data set. Both methods are based on the analysis of the covariance matrix of the initial dimensions and, although in practice the results of the two approaches use to be very similar, there are reasons to distinguish one from the other (Fabrigar *et al.* 1999).

Firstly, FA is based on the *common factor model* proposed by Thurstone[33] (1931; 1947), according to which each variable x_i is a linear combination of a set of common factors $\{f_1, f_2, ... f_n\}$ and a unique –or specific– factor u_i:

$$x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots + \lambda_{ik} f_k + u_i$$

Or, in matrix terms:

$$\mathbf{X} = \Lambda \mathbf{F} + \mathbf{U}$$

where λ_{ij} are the factor weights –or loadings–, and Λ is the weight matrix[34].

[31] Because of this, these techniques have been widely applied in intelligence research and cognitive science (Barlow 1959; Edelman and Intrator 1997; Edelman 1998; Jensen 1998; Tenenbaum *et al.* 2000; Sternberg 2005).

[32] The reason why a reduced-dimensional space is enough to recognize stimuli originally located in a much larger dimensional space is the great redundancy present in the external input.

[33] Even though Thurstone was the first who used the expression *factor analysis*, Spearman (1904; 1927) had previously used the term *factor*.

The difference between a *common factor* and a *unique factor* is that the former can influence over more than one observed variable, while the later influence only over one particular variable. The aim of the common factor model is to describe the variability of the observed variables by estimating the relationships between the common factors and the observed variables. Or, in other words, FA distinguishes between *common variability* –associated to several variables– and *unique variability* –associated to only one variable–, and tries to find a new set of dimensions which is able to describe as much common variability as possible.

There exist two main fitting methods for estimating the factor loadings:

— *Maximum likelihood*: this method assumes that data come from a multivariate normal distribution, and estimates the factor loadings by maximizing the likelihood function of the *k*-factor model, by means of an iterative process.

— *Principal axis factoring*: this approach applies a PCA to the reduced covariance matrix, which is equal to the covariance matrix except for its diagonal values, which have been replaced by the communalities[35] of the observed variables.

Because the principal axis method does not assume normally distributed data –which cannot be guaranteed for the case of perceptual inputs–, this could be a reason for preferring the latter method in the case of concept acquisition.

* * *

By contrast, PCA does not discern between common variability and unique variability. This model, originally proposed by Pearson (1901) and Hotelling (1933), defines each observed variable as a linear function of a set of –linearly– uncorrelated variables $\{z_1, z_2, \ldots z_n\}$, known as *principal components*. Each principal component z_j is defined as the linear combination of the observed variables with maximal variance, subject to being uncorrelated with all the previously obtained principal components $z_1, z_2, \ldots z_{j\text{-}1}$. On this basis, it can be proved that the vector of principal components **Z** is related to the vector of observed variables **X** by means of the matrix of eigenvectors **A** –of the covariance matrix of **X**–:

$$\mathbf{Z} = \mathbf{XA}$$

Therefore, the calculation of the principal components is equivalent to applying an orthogonal transformation **A** to the observed variables **X** (in order to obtain a set of uncorrelated dimensions **Z**). Lastly, the dimensionality reduction takes place when the first *k* principal components $\mathbf{Z}_k$ are chosen:

$$\mathbf{Z}_k = \mathbf{XA}_k$$

[34] For a comprehensible introduction to modern FA, see Harman (1967), Lawley and Maxwell (1971), and Yates (1987).

[35] The communality of a variable x_i is the part of its variance that is explained by the common factors **F**.

where $\mathbf{A}_k$ is the matrix $\mathbf{A}$, truncated to the first k eigenvectors –sorted by decreasing order–[36].

In both cases (i.e., FA and PCA), the statistical processes produce a number k of factors, or components, lower than the number n of original variables. This notwithstanding, since FA tries to maximize the captured common variability –in contrast to PCA, which tries to maximize the total variance–, it seems that FA is a more suitable candidate in order to identify the features shared by the members of a given category.

Lastly, in PCA, but above all in FA, the set of reduced dimensions can be, and usually is, rotated, in order to find a more optimal result. That better solution would be one in which the coefficients that relate the reduced factors to the original dimensions are as simple as possible (i.e., either very close to zero, or very far from zero). This has a significant influence on the resultant model since: (a) a smaller number of *rotated* factors explain a larger amount of the variability in the original data; and (b) the resulting factors are more easily interpretable.

Application case

Let us consider a visual perceptual input. In this case, the external stimulus might be constituted by a two-dimensional $m \times m$ array of intensity values, one for each retinal point. (In a similar way, if an auditory perception had been considered, the perceptual stimulus could be identified with a one-dimensional array of n values with frequency information.)

If, for instance, the visual input of the acquisition system were constituted by a set of images of faces, then a dimensional reduction could be carried out over those images. Turk and Pentland (1991) performed this kind of research, and analyzed a group of face images by means of PCA, obtaining a set of k best eigenvectors (principal components) from a total of m^2 eigenvectors. Those k best eigenvectors can be thought of as the constitutive features of the set of face images[37].

And because this sort of analysis may be carried out on any set of images –or, more generally, on any kind of input–, in order to produce a reduced set of variables, then the application of successive dimensional reductions over the chosen input will lead to the progressive identification of its constitutive factors –properties or features–.

Consider, for example, the data set shown in Table 6.1. A principal component analysis carried out over those data would produce a set of factors such that the first four of them explain the 81.5% of the variability –i.e., variance– in the input data (see Table 6.2). This means that it is possible to explain more than 80% of the variability in the data, and reduce by 73% the information required for it.

[36] For an introduction and review of PCA in modern data analysis, see Jackson (1991), Jolliffe (2002).

[37] For a review of many other applications of both FA and PCA in the fields of cognition and perception, see Carroll (1993).

Automatic feature extraction

In consequence, a dimensionality reduction system working over a set of raw data may automatically produce new features in the absence of a previous conceptual structure. This is so because those data analysis methods (i.e., FA and PCA), merely operate on numbers (that is, the values contained in the input data), from which they generate a set of resulting reduced factors –on the basis of their ability to remove redundancy from the input data–. Therefore, it is fair to say that these approaches are feature extractors, able to transform a high-dimensional input space into a low-dimensional one constituted by a reduced set of non-redundant variables –factors or components– relatively invariant to transformations, and which might be identified with the features –or most basic elements– of our conceptual systems.

Obs.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	9	10	8	9	9	10	9	8	10	9	9	9	9	8	9
2	5	3	8	3	3	1	4	1	2	6	3	3	2	4	3
3	6	6	6	3	8	3	8	4	4	8	8	6	2	4	7
4	2	10	3	9	10	5	7	9	5	10	8	4	10	0	8
5	8	9	9	9	10	5	10	10	10	5	9	9	8	8	8
6	0	2	10	2	5	0	6	0	0	10	0	0	0	1	0
7	4	6	10	5	8	3	10	5	5	7	9	2	6	9	7
8	8	7	10	6	8	2	8	5	9	7	9	4	8	8	9
9	3	8	7	8	10	2	7	8	7	10	10	4	10	4	10
10	8	5	10	5	8	2	5	3	3	4	4	5	4	7	3
11	6	5	10	6	8	9	7	5	6	7	7	7	6	6	8
12	6	4	8	2	6	2	8	2	3	5	6	9	6	5	5
13	5	5	8	4	5	3	8	2	6	7	4	7	5	8	6
14	0	0	2	2	2	10	7	0	10	8	1	10	0	0	3
15	5	3	10	5	7	2	3	3	2	5	9	3	5	7	7
16	3	10	7	10	10	3	7	10	10	10	10	4	9	2	10
17	3	1	5	2	2	6	4	3	8	9	4	3	3	8	3
18	7	3	10	3	9	1	6	3	3	5	4	4	2	6	2
19	6	5	9	4	5	4	9	4	8	8	4	6	7	10	4
20	4	3	10	5	5	7	8	2	6	8	4	4	6	7	6
21	0	0	0	0	1	10	3	0	10	8	1	10	0	0	0
22	8	9	9	9	9	4	8	7	10	3	8	7	6	6	8
23	10	10	10	10	9	10	6	10	10	9	10	10	10	10	10
24	5	7	10	6	8	3	8	5	8	6	8	3	8	8	8
25	5	4	0	4	3	0	3	0	0	4	0	4	0	3	0
26	10	9	10	9	9	3	10	10	8	7	9	7	9	9	10
27	10	8	10	10	9	3	8	10	8	7	10	9	8	8	10
28	5	4	9	4	4	2	9	3	4	7	5	6	4	7	4
29	6	7	8	8	8	8	9	9	10	6	9	6	8	7	6
30	3	9	8	3	10	1	9	9	2	6	8	4	5	4	7
31	5	6	5	2	7	4	7	4	4	8	8	6	3	4	8
32	7	8	8	7	9	6	7	6	8	7	5	8	6	7	6
33	4	7	10	5	7	1	8	3	4	10	9	9	9	10	9
34	0	2	10	2	5	0	7	0	0	10	0	0	0	3	0

35	7	5	9	5	4	8	8	5	7	8	4	6	8	8	8
36	5	3	5	3	5	0	3	0	0	5	3	6	0	3	0
37	5	2	9	0	3	1	7	0	3	4	0	6	1	3	3
38	4	5	10	5	7	1	6	2	5	10	7	9	8	9	7
39	7	9	8	8	8	3	7	8	10	2	7	6	5	5	7
40	5	3	10	2	7	2	3	3	2	5	9	2	4	7	6
41	8	10	8	8	8	10	9	8	10	8	8	9	9	8	8
42	5	10	7	2	9	1	9	9	2	6	9	4	5	6	8
43	10	7	10	9	8	2	8	10	8	7	10	9	9	10	9
44	5	6	8	2	6	5	9	2	6	8	3	8	7	9	6
45	6	5	9	4	6	8	6	2	5	8	5	5	7	5	8
46	8	9	8	8	8	9	9	5	10	7	8	9	8	8	8
47	10	8	10	10	9	10	6	10	10	9	10	10	10	10	10
48	6	10	8	8	9	8	7	6	5	7	8	7	8	6	8

Table 6.1. Principal component analysis: example data set, with 48 observations characterized by means of 15 observed variables x_i, which vary from 0 to 10.

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	z_{13}	z_{14}	z_{15}
eigenvalue	7.51	2.06	1.46	1.20	0.74	0.49	0.35	0.31	0.26	0.18	0.15	0.10	0.09	0.06	0.04
% of var.	50.1	13.7	9.7	8.0	4.9	3.3	2.3	2.1	1.7	1.3	1.0	0.7	0.6	0.4	0.2
% accum.	50.1	63.8	73.5	81.5	86.4	89.7	92.0	94.1	95.8	97.1	98.1	98.8	99.4	99.8	100.0

Table 6.2. Principal component analysis: variability explained by each factor –or principal component–: the first row shows the eigenvalues of the covariance matrix of **X**; the second row presents the variability in the data explained by each component z_i; and the third row represents the accumulated variability explained by z_j and all the previous factors z_i (with $i < j$).

This kind of techniques extract the relevant information contained in the original variables, by capturing as much input data variability as possible in a set of reduced dimensions[38]. Thence, they are an approach based merely on experiential input, and able to produce new features (i.e., the resulting factors or principal components) without resorting to previously existing conceptual or perceptual elements[39]. By virtue of all this, di-

[38] Although in order to provide an explanation of concept acquisition the only mandatory processes are those of pattern recognition, the seemingly optional character of dimensionality reduction shifted to compulsory when other questions are taken into account. On the one hand, the reduction in the number of dimensions was useful from a *computational* point of view since it largely reduced the size of the original raw data, which made their processing and storage into less resource demanding. From a *cognitive* point of view, a lower number of dimensions increased the efficiency of the pattern identification processes working on them. Therefore, from here on I will consider dimensionality reduction as *de facto* mandatory.

[39] Indeed, my proposal should be viewed as an *existence proof* of an empiricist learning mechanism which refutes nativist arguments against the acquisition of primitive concepts; and not as an advocacy of the psychological plausibility of the particular proposed model. In fact, although I have a strong belief about the general structure of the described approach (that is, a kind of iterative learning system,

mensional reduction provides a simple and natural answer to the selection problem –aforementioned in section 2.3.1–, according to which concept theories need to explain how the set of relevant properties is chosen. On my view the answer is that the relevant dimensions are those which rule out as much redundant information as possible. Additionally, in an approach like the one here proposed, they can be determined in an automatic and unsupervised way.

All in all, this is the first and main cognitive module by means of which an answer to the circularity threat in concept empiricism may be provided. Consequently, it may be said that *dimensionality reduction is the place where the features come from*.

6.4.4 Pattern recognition

Up to this point the proposed approach has merely produced a reduced dimensional space, within which the original positions of objects can be relocated, through a transformation of the original variables into the reduced ones. However, that is of no use for classifying those objects into categories relevant for the subject. By virtue of this, a pattern recognition system is required. In this case the cognitive system must produce a computationally-efficient classification function, able to robustly map objects into categories useful for the subject. Obviously, those categories will not be known before the end of the learning process or, in other words, the subject's mind has no *a priori* knowledge about the pursued categories.

Difference between identifying-patterns and acquiring-concepts

Once a category has been learned, the pattern recognition module will apply its associated classification function in order to produce a positive response in the presence of the right stimulus. And, because the input of the pattern recognition step could be constituted both by (dimensionally reduced) sensorial data and by patterns previously acquired, the output resulting from that process may be identified: (a) either with the most basic elements of our cognitive system –if the input of the process were merely composed by sensorial data–; (b) or with complex concepts, wholly or partly constituted by other simpler concepts.

This second type of cognitive ability (i.e., *pattern identification*) articulates what is commonly referred to as concept acquisition. Independently of how it is deployed in the mind, the main goal of a pattern recognition process is to identify regularities and patterns in the data input (either in the original ones, or in those resulting from the dimensionality reduction step). The most basic output of this kind of process would be a discriminant function (or, alternatively, the parameters of a discriminant mechanism[40]) able to single out similar stimulus experienced in comparable circumstances.

where a dimensional reduction is followed by a pattern recognition stage), I am far more circumspect in regard to the MVA techniques (i.e., FA/PCA and cluster analysis) by means of which I have characterized dimensional reduction and pattern recognition –respectively–, since these two general-purpose modules could be articulated otherwise –as I have repeatedly said throughout all this chapter–.

[40] Although both a *discriminant function* and a *parameterized discriminant mechanism* (PDM) can play the very same distinguishing role, they are different from an architectural point of view. On the one

Ultimately, in order to be an explanation of how the general formation of concepts happens, the learning process must be unsupervised, in the same way that the human concept acquisition system seems to work. This opposes to the working of supervised learning systems, where patterns are acquired from sets of labeled examples and learning is guided by prior knowledge. By contrast, when learning is unsupervised, data input is unlabeled and there is no *a priori* guidance of the learning process. This is the reason why, on my view, it is appropriate an approach based on a purely statistical dimensional reduction, followed by a "blind" search of patterns, whose validation happens at the end of each iteration (of the process) –in terms of the temporal associations successfully predicted from all those present in the stored historical data–.

Consequently, the stage of pattern recognition should not be identified with the whole concept acquisition process, although it might be tempting to do so, as many times it is done. Maybe the reason of this misunderstanding is that such identification is valid in the sense that the obtained pattern –either in the form of a discriminant function or in the form of a PDM– is what determines the classification of an object under a certain category. However, even though the categorization of something as a particular concept happens through the application of a previously obtained pattern, things are different when we consider how a concept is acquired. In the latter case, the recognition of a pattern is not equivalent to the acquisition of a concept, because it is possible that many provisional patterns have to be obtained (and that many dimensional reductions have to be tested) before one pattern is definitively chosen and assigned to a given category. That is, concepts result from the whole iterative process, and not only from one execution of its second (pattern recognition) stage.

Cluster analysis

The earliest precedents of cluster analysis (CA) techniques are found in the field of anthropology (Czekanowski 1911; Driver and Kroeber 1932), although those ideas were soon applied in psychology (Zubin 1938; Tryon 1939; Cattell 1943)[41]. In a nutshell, cluster analysis refers to a set of multivariate statistical models whose goal is to group objects with similar properties into the same clusters, and dissimilar objects into different clusters, being a well-established paradigm both in supervised and unsupervised learning[42]. From here on I will mainly focus on the unsupervised approaches to cluster analysis. And although there are multiple clustering algorithms, which could be differently

hand, it may be said that a function *f* is discriminant if the equation $f(x)=0$ defines a decision surface dividing the space into regions associated to the considered patterns. On the other hand, a PDM could be described as a decision mechanism –which might be innately given to the subject– (e.g., a system able to determine the distances between different points within a conceptual hyperspace), together with a set of parameters that determine how the PDM works.

[41] For a brief summary of the history of cluster analysis, see Wilmink and Uytterschaut (1984). For a review of contemporary cluster analysis, see Bailey (1975), Blashfield and Aldenderfer (1988), Grabmeier and Rudolph (2002), Jain (2010), and Everitt *et al.* (2011).

[42] Sometimes *unsupervised* clustering is also called *intrinsic* classification, since no category labels are used in order to carry out the partitions of the objects. In *supervised* clustering labels are considered when partitioning data, and classification is called *extrinsic* (Jain and Dubes 1988, pp. 56-57).

parameterized, all of them are iterative optimization mechanisms whose parameters and preprocessed data input often need to be revised[43]. In order to assess the quality of the resulting groups, two kinds of approaches may be applied: (a) *internal*, when the evaluation is based on the same data used by the cluster analysis; (b) *external*, when groups are evaluated with information not used by the clustering.

Cluster analyses take as starting point a set of *p* observed variables about the set of *n* considered objects. Such information may be represented by a matrix **X** whose rows stand for the objects $\mathbf{x}^i = (x_1^i, x_2^i, \ldots, x_p^i)$ (where $\mathbf{x}^i$ represents the *i*-th object, with components x_j^i), and whose columns stand for the variables $\mathbf{x}_j = (x_j^1, x_j^2, \ldots, x_j^n)$ (that is, the values of the *p* variables for each of the *n* objects).

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \cdots & x_p^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_p^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^n \end{pmatrix} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$$

Additionally, cluster analysis can be viewed as a geometrical model of similarity, where distances are determined as was said in section 4.1. Nevertheless, in cluster analysis it is necessary to take into account, not only distances between particular objects, but also distances between-group and distances within-group:

— *Within-cluster distance*: it is determined as a measure (i.e., sum, average, maximum, or minimum) over one of the three following sets: (a) distances between all the pairs of points in the cluster; (b) distances between the centroid and all the points of the group; or (c) distances between the medoid and all the points of the cluster[44].

— *Between-cluster distance* (for two clusters): it can be characterized through different approaches: (a) distance between the centroids / medoids of the considered clusters; (b) average distance between all the pairs of objects that may be formed with the elements of the considered groups; or (c) distance between the closest or furthest points of those groups.

And, because the aim of cluster analysis is to classify objects into homogeneous groups –and different to other produced clusters–, that objective is achieved when within-cluster distances are minimized, and between-cluster distances are maximized.

[43] By contrast, in my approach parameter calibration is carried out by the third stage of the iterative learning process (in which the model is evaluated and readjusted), while the preprocessing of the input data would correspond with the dimensionality reduction stage.

[44] The *centroid* of a cluster is the mean position of all the points in the group, which leads to a point whose dissimilarity to all the points of the cluster is minimal. The notion of *medoid* is analogous to that of centroid, with the difference that it must be a point belonging to the cluster (i.e., the medoid of a cluster is the point of the cluster with minimal average dissimilarity to all the rest of points of that group).

Nonetheless, by virtue of their heuristic character, there is a wide variety of cluster analysis techniques. In particular, it is possible to distinguish the following three main approaches (Bailey 1975; Aldenderfer and Blashfield 1984):

[A] *Hierarchical methods*: this kind of approach can be traced back –at least– to the works of Cox (1957), Fisher (1958), and Sokal and Sneath (1963), and tries to accomplish the goal of CA by producing a sequence of partitions, through one of the following procedures (Gordon 1987, 1999): (i) *agglomerative algorithms*, which begin with *n* singleton groups, with *n* equal to the number of considered objects, and in each step combine the two closest groups into one only new cluster; (ii) *divisive algorithms*, which take as starting point a single class, and in each step divide it (or another already produced class) –successively– into two new clusters, until singleton groups are achieved; (iii) *incremental* (or *constructive*) *algorithms*, which work sequentially adding new objects to the analysis.

In all these cases, hierarchical clustering produces a series –or tree– of partitions that may be represented by means of a *dendrogram*, where the height of the lines represents between-cluster distances –or, inter-group dissimilarities– (see Fig. 6.2). Lastly, the partition trees produced by hierarchical clustering represent disjoint *final* clusters, and the *intermediate* clusters wholly contain the subgroups under them. (This is the reason why they are called *hierarchical methods*.)

[B] *Nonhierarchical methods* (or *partitional clustering methods*): in contrast to hierarchical approaches –which produce partition trees–, the result of nonhierarchical clustering is a set (not a tree) of nonoverlapping clusters[45]. The best-known nonhierarchical approach are *k*-means algorithms, which have its origins in the works of Steinhaus (1956) and Lloyd (1957), even though the name "*k*-means" was firstly used by MacQueen (1967)[46]. And, even though many algorithms for *k*-means have been proposed, all of them use to begin with an initial set of centers –or initial "means"[47]– of the groups, which are moved from one step to the next trying to maximize between-cluster distances and/or to minimize within-cluster dissimilarities. For instance, Lloyd's (standard) algorithm tries to minimize the following within-cluster sum of squares:

$$\sum_{j=1}^{k} \sum_{\mathbf{x}^i \in \mathrm{S}_j} || \mathbf{x}^i - \boldsymbol{\mu}_j ||^2$$

where *k* is the number of tentative subgroups –or clusters–, $\mathbf{x}^i$ represents the *i*-th object, S_j represents the *j*-th subgroup, and $\boldsymbol{\mu}_j$ is the centroid of S_j.

[45] Other difference between hierarchical and nonhierarchical approaches is that partitional clustering tries to determine all the clusters at the same time –and not step-by-step, one after another–.

[46] For a review of the history of *k*-means algorithms see Bock (2007) and Jain (2010).

[47] These initial means –or initial seeds– could be innate information used by the pattern recognition stage of an empiricist approach to concept learning (as was argued in section 6.2.2 above).

Nonhierarchical methods –in general– and *k*-means algorithms –in particular– have less computational complexity, and are more efficient (i.e., faster) than hierarchical methods, so they are more appropriate for analyzing huge amounts of data like those present in the problem of concept acquisition.

[C] *Other clustering approaches*: in addition to hierarchical and nonhierarchical clustering, there are other possible procedures that can lead to the classification of objects into homogeneous groups: (i) *overlapping clustering*, where objects may be included in more than a cluster at the same time, as happens in models of additive clustering (Shepard and Arabie 1979; Mirkin 1987); (ii) *fuzzy clustering*, where the membership of an object $\mathbf{x}^i$ in a cluster S_j is a matter of degree, and –due to this partial character– it can take any real number between 0 (i.e., no membership) and 1 (i.e., full membership) (Dunn 1973; Bezdek 1981).

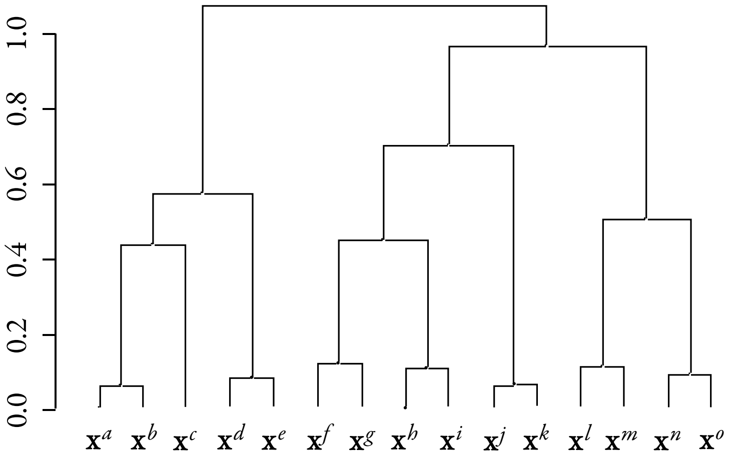

Fig. 6.2. Example of dendrogram for a hierarchical clustering. The final objects $\mathbf{x}^i$ are represented by the leaves of the tree in the lower part of the graph. The vertical axis is proportional to the distances between the daughter clusters (i.e., the between-cluster dissimilarities of the underlying nodes).

Automatic pattern identification

Thus, a pattern recognition system based on CA can produce the centroids which –in a similarity-based theory of concepts articulated by means of prototypes– were identified with the notion of *stored concept*, and from which *instantiated concepts* were generated (see chapter 5 of this thesis).

However, any clustering algorithm can produce a variety of resulting groupings, so it is necessary to determine which of those groupings is more appropriate in the subject's context and circumstances. And, although this is the topic of section 6.4.5 below, something has to be said at this point regarding how the groups resulting from CA are evaluated. Firstly, as described in footnote 42 of this chapter, depending on whether the learning process is supervised or unsupervised, the evaluation of classification will be extrinsic or intrinsic, respectively. I have also said that the pattern identification module must be viewed as an unsupervised process, so I will focus on how the groups produced by CA can be intrinsically evaluated. Secondly, although intrinsic classification resorts only to the information on distances in order to carry out the CA, the resulting groups may be evaluated in two different ways:

- *Only on the basis of distances*: in this case the quality of partitions is evaluated by means of an appropriate homogeneity measure given –by instance– in terms of the within-cluster sum of squares, and the final number *k* of clusters[48]. Because different initial seeds and/or different numbers of clusters will produce distinct partitions, such a kind of homogeneity measure would allow to *internally* compare each grouping result with the others. Nonetheless, one disadvantage of internal evaluation is that a high score on an internal measure does not entail that the resulting clusters will be practically useful/effective for the subject[49]. Or, in other words, even though this approach tests the internal quality of the model, it does not guarantee the identification of patterns relevant for the subject. That is the reason why a second kind of evaluation is needed.
- *On the basis of both distances and a priori labels/information*: if the evaluation is based on data which are not used for the clustering, it is usually called *external* evaluation. That kind of external data includes *a priori* category labels that preclassify some groups of objects, and which can be the starting point for measuring how close a particular clustering is to a predetermined set of classes. Unfortunately, a set of *a priori* labels is not something that is available to guide us in the acquisition of new concepts. By virtue of this, an additional kind of evaluation and readjustment of the model is demanded, and that is the subject of the next section.

6.4.5 Evaluation and readjustment of the learning process

Now I will show how the two aforementioned learning abilities may work together in a unique and integrated cognitive process, which explains how the acquisition of concepts –both primitive and complex– takes place. Here my first and main thesis is that the global learning process must be conceived as an iterative one, stepping sequentially through the three following stages: (i) dimensionality reduction, (ii) pattern identification, and (iii) evaluation of results in terms of the predictive power of the final patterns, and global readjustment of the iterative process.

Nevertheless, why should an iterative process be necessary? It might be argued that the processes of dimensional reduction and pattern recognition are more than enough to explain how concept acquisition happens and, consequently, that the introduction of a three-step iterative system is an unnecessary artifice. On my view, there are two mains reasons to support the thesis that an iterative process is required. The first is that nothing guarantees that the factors resulting from a particular dimensionality reduction are the

[48] In this case, the homogeneity value would be inversely proportional, both to the within-cluster sum of squares, and to the number of final clusters.

[49] For a review of internal evaluation methods of CA, see Manning *et al.* (2008, chs. 16-17).

best for the subsequent pattern recognition stage[50]. The second is that there is no guarantee that the patterns obtained in a certain realization of the learning process are the most predictive possible ones from the available sensorial data. The question at this point is how the third stage –i.e., evaluation and readjustment of the model– of the acquisition process could work.

With regard to the evaluation of results, both the reduced factors and the obtained patterns have to be judged in order to determine their appropriateness in practical and cognitive terms. If both of them are acceptable, then the reduced dimensions –up to this point, with tentative character– will be confirmed as features provisionally valid –or effective–, and the same holds of the patterns resulting from CA, which will be confirmed as concepts provisionally[51] valid at that point. By contrast, if results are not suitable, the reduced dimensions and identified patterns remain tentative, and the iterative process returns to stages (1) or (2) (i.e., the dimensionality reduction or pattern recognition stages[52], respectively (see Fig. 6.1)), until a more apt solution is accomplish or the iterative process ends –once considered any possible alternative, or once a certain number of iterations is surpassed–.

But, how factors and patterns can be internally evaluated (without resorting to any set of *a priori* information)? The idea is to judge the appropriateness of results on the basis of their ability to successfully predict temporally ordered associations present in the original data. Indeed, if data are stored in a way that reflects their simultaneity and temporal order (i.e., if data are *temporally structured*), as was assumed in section 6.4.2, then it is possible to search for temporal associations[53] between the previously identified factors and/or patterns.

This approach has two main advantages: (a) on the one hand, the only considered information are the original input data, which allows an internal evaluation of results –in line with the claimed kind of (unsupervised) learning–; (b) on the other hand, diachronic associations are something that has not been taken into account up to this point of the process, so it is information genuinely new that may be conveniently used as a benchmark for the obtained factors and patterns. On those bases, the evaluation process might run as follows:

(1) Calculation of the obtained factors and patterns for different moments in a particular time interval.

[50] Even worse, nothing guarantees that the reduced factors –or (tentative) obtained features– will be useful for higher-level cognitive processes, like those of the pattern recognition stage (Mozer 1994).

[51] *Provisionally* in the sense of valid from the end of the iterative learning process till the next execution of it, which will happen on the basis of new –and old– perceptual data available to the subject.

[52] Depending on whether the non-suitable character of the achieved result is due: (i) to the reduced factors, (ii) to the recognized patters, or (iii) to both of them.

[53] These temporal associations should not be confused with the patterns that resulted from the second stage of the proposed model. The difference is that, while the regularities identified by the pattern recognition module are *synchronic*, since they were produced from one set –or multiple sets– of transversal information; the patterns now considered are *diachronic*, being produced from longitudinal information.

(2) Search of temporal associations between those factors and patterns; or, in other words, search of recurrent relations of precedence/antecedence between the factors and patterns previously identified[54].

(3) Computation of the objective function, which serves as an appropriateness –or fitness– metric for the achieved solution. This objective function is a quantitative measure which acts as an evaluation criterion of the success of the learning process, not only on the basis of the number of obtained associations, their frequencies, and degree of complexity[55]; but also on the basis of the subject's needs, goals, and interests, which may be either innately given (e.g., those associated to biological functions) or learned/acquired.

Lastly, with regard to how the proposed iterative process can evolve in function of the adequacy of results, two kinds of readjustment may happen:

— *A total or partial change of the dimensionality reduction*: this could be due to different reasons: (a) problems with the factors resulting from the dimensional reduction, if some of them played little or no relevant role in the characterization of patterns; (b) problems because the reduced factors did not produce a predictive enough set of patters –in terms of the above-mentioned temporal associations–; (c) a new search of solutions, in order to determine if other alternative dimensional reductions can lead to the recognition of more predictive patterns.

— *New search for relevant patterns*: in a process that could retain –or not– some of the patterns identified in previous iterations, from the same or from a different dimensional reduction, depending on whether the first adjustment was applied or not. The main causes of this second adjustment are the same as the last two reasons pointed out for the first adjustment –that is, reasons (b) and (c) of the previous point–.

As a result of those readjustments the iterative process may: (i) return to the dimensionality reduction stage, in order to obtain new reduced factors[56]; (ii) return directly to the stage of pattern recognition, if no change is needed in the resultant factors; or (iii)

[54] The identified temporal associations might give rise to –or be interpreted in terms of– causal Bayes nets. For accessible expositions on causal Bayesian networks, see (Glymour and Cooper 1999; Glymour 2001; Sloman 2005; Gopnik and Schulz 2007).

[55] When contrasting different solutions, the greater the number of associations, the better the solution; the lower the degree of complexity, the better the solution.

[56] This first kind of feedback is in line with (a) the *functionality principle* (Schyns *et al.* 1998), according to which the learned categorizations (i.e., the regularities, or tentative categories, identified by the pattern recognition stage) should have an influence on the set of extracted factors; and also with (b) the substantial evidence of top-down effects (Rolls 2008; Gilbert and Li 2013; Zhang *et al.* 2014), according to which lower brain regions may be modulated by higher regions (where *lower* and *higher* must be understood in terms of their place in the processing hierarchy). For other examples of the influence of high-level cognition over lower level conceptual / perceptual processes, see Goldstone (1994; 1995), Schyns and Rodet (1997), Schyns *et al.* (1998), and Goldstone *et al.* (2015).

finish, if a termination condition happens[57]. In this third case, the best solution reached up to this moment is confirmed, and moved from tentative to definitely valid for that realization of the learning process. From then on, the confirmation of the reduced factors and classification patterns leads to their effective application by the subject's cognitive system until the next execution of the learning process.

6.4.6 Final remarks

An iterative process like the one described above has significant implications for the underlying cognitive system. Firstly, the external raw data are not the most basic constituents of the conceptual system, since that they cannot even be persistently stored beyond their daily processing and analysis. Second, those most basic elements of concepts are the factors that result from the dimensional reduction, so –against Fodor's arguments– they are not innately determined. And, inasmuch as the constitutive elements of concepts are not an input, but an output of the learning process, the proposed model proves that it is possible to provide an account of concept acquisition –both of concepts and of their primitive constituents–, by means of a general-purpose learning mechanism (in line with empiricist thesis) without falling into circularity.

My proposal shows that primitive concepts (i.e., the most basic elements of concepts) can be the mere result of a process of redundancy reduction, in which no *a priori* information about the pursued categories is given. Then tentative patterns will be recognized on the basis of the factors resultant from the dimensionality reduction stage, and both of them (i.e., factors and patterns) will be evaluated through their predictive power –measured in terms of the temporal associations identified from those factors and patterns–. In this type of approach, the precedence assumption is not needed by the model. As a consequence, its explanation of how primitive concepts are acquired is free from the threat of circularity.

Additionally, my perspective also enjoys other significant advantages over other alternative approaches. Firstly, the proposed model takes as starting point perceptual information, which is available to the subject as sensory input data; in contrast to other views where the data required by the analyses are not available to the subject as an input of the learning process –e.g., the approaches based on multidimensional scaling, MDS, which need as an input the (dis)similarity matrix of the considered objects (see sections 3.3.2 and 3.3.3)–. Secondly, even though my proposal is an empiricist acquisition model, it allows to accept innate elements –in the form of the initial seeds of the CA– which do not constrict the empiricist character of the model, nor prevent its adaptability to changing environments.

In the third place, an approach like the one here proposed constitutes a response to the *selection problem* which affected any similarity-based theory of concepts (see section 2.3.1). Indeed, under this kind of view the selection problem is a pseudo-problem. As

[57] There can be multiple termination conditions, but the three main ones are the following: (1) that the processing time –or biological devoted slot– is over; (2) that every possible readjustment of the dimensional reduction and pattern recognition stages had been tested; and (3) that the marginal improvements of the last *n* iterations did not exceed a certain threshold.

repeatedly said by psychologists, our concepts only express some of the many properties typically present in the members of each category. However, that is only a problem if a concept is viewed as something to be *a posteriori* characterized in terms of a subset of the properties common to all or most of its members. Or, in other words, the selection of properties is only a problem if concepts are considered as something given (e.g., as categories present in the world), and we wondered about the set of features that binds their members together. However, from an empiricist point of view, doing that is to put the cart before the horse, because concepts are not something that preexist their constitutive properties. Indeed, in this kind of approach the most basic constituents of concepts are the result of a first process of dimensionality reduction; and only then a pattern recognition process is run over the set of (tentative) reduced factors, whose result will be a set of provisional categories. Thus, properties are not something to be looked for once a certain category has been obtained, but the elements from which that category is produced.

6.5. Conclusions

Throughout this chapter I have claimed that one of the main problems of empiricist theories of concepts is to explain how the most basic constituents of concepts may be acquired, without resorting to a preexisting innate repertoire. In fact, that was the origin of Fodor's nativist critiques against concept empiricism, and of many other reformulations of Fodor's argument.

I have also shown that the crucial problem of concept empiricism is not due to the Fodor's or Carey's premise that all learning mechanisms may be reduced to hypothesis formation and testing, but to an assumption tacitly present in most of empiricist and nativist learning models, to which I gave the name of *precedence assumption*. According to it, the perceptual or conceptual constituents of a concept have to be available as an input of the learning process that leads to the acquisition of such a concept. My point was that the acceptance of the precedence assumption is a mistake for the empiricist, because no model built on it may explain the acquisition of general concepts in a non-circular way.

However, I have argued that the precedence assumption is an unfounded premise, because the constitutive elements of a concept may result from the same learning process by virtue of which that concept is acquired. In that case, it would be enough that the most basic elements of concepts were ready at the end of the acquisition process, and not from its beginning, so the precedence assumption is not needed, which prevents the threat of circularity when explaining how concepts are acquired.

Finally, in the last section of the chapter I have described a learning system which, on the basis of a three-step iterative approach, was able to produce concepts and their constitutive properties as result of the same execution of the learning process. Such a system was based on two general-purpose modules, namely, dimensionality reduction and pattern identification, followed by a final stage that evaluated the obtained results and readjusted the model.

On the one hand, the dimensionality reduction module was able to produce new relevant (and reduced) factors which ruled out as much redundant information as possible. That is, the new tentative features / properties, in general, and the most basic constituents of concepts, in particular, may be produced in an automatic and unsupervised way, as those which minimize redundant information while retaining as much variability (of the original data) as possible. On the other hand, the pattern recognition module will find regularities in the set of reduced factors, and as a result will generate –for an approach based on cluster analysis– the centroids that were identified with the notion of *stored concept*, and from which *instantiated concepts* are produced.

In consequence, once the precedence assumption is surpassed by this kind of model, the circularity threat disappears, and innateness is no more a necessary condition in order to explain the acquisition of primitive concepts, which refutes Fodor's claim that empiricists have to accept the innateness of a primitive conceptual repertoire.

This page intentionally left blank

Conclusions

But the fact that Utopia is a long way off does not mean that daily life should come to a screeching halt. There is plenty for us to investigate, in our sloppy and impressionistic fashion, and there are plenty of real results to be obtained. –
Hilary Putnam (1970, p. 201)

The main goal of this thesis has been to show that concept empiricism has a way out of the nativist critiques against the thesis that primitive concepts can be learned without relying on a preexisting innate set of concepts. More precisely, the final aim was to provide an empiricist model which explained the acquisition of primitive concepts in a non-circular way. The result, on my view, is more than acceptable.

The first three chapters were the introductory part. First, in chapter 1 I described the different theories on the nature, origin, internal structure, and context-dependence of concepts. Then, in chapter 2 I examined the principal theories on the structure of concepts, which were mainly based either on definitions, similarities or explanations. Lastly, chapter 3 was devoted to the idea of similarity and the various similarity-based approaches to the structure of concepts. In each of those chapters I made explicit my assumptions regarding all those issues. More in particular, I opted for an empiricist-contextualist approach articulated by means of a geometric similarity-based model of the prototype theory of concepts.

With that framework in mind, in chapter 4 I discussed the notion of conceptual space, as a way of characterizing concepts and knowledge. There I focused on Gärdenfors' theory of conceptual spaces, and I showed that the convexity requirement on the geometry of (conceptual) regions is both unnecessary and problematic. On this basis I concluded that, if the convexity constraint is given up to, then Gärdenfors' conceptual spaces can be reduced to a (geometric) contextualist particularization of the prototype theory of concepts.

The last two chapters were devoted to develop my main proposals in this doctoral thesis. First, in chapter 5 I investigated how the approach assumed by me in this work (i.e., a contextualist similarity-based conceptual space theory) can join together virtues both from the invariantist and from the contextualist sides. In that chapter, I showed that two different notions of concepts must be distinguished, which were associated with two distinct facets in the life cycle of a concept (i.e., storage and instantiation): (a) stored concepts contained the information needed to be –persistently– maintained by the mind of a concept for the subsequent instantiation of it; (b) instantiated concepts were the result of the cognitive processes which applied concepts in categorizations, inferences, etc. As a consequence, stored concepts could account for significant invariantist phenomena, while instantiated concepts were able to explain aspects typically explained by contextualism. Finally, I argued that, if concepts are thought to be context-dependent, then instantiated concepts lack minimal persistence and, consequently, cannot be a representation of their associated categories.

Then, chapter 6 researched whether the acquisition of primitive concepts could be explained by means of an empiricist approach, without relying on an innate repertoire of elements. There I proved that the uppermost nativist arguments against the acquisition of primitive concepts depend on the precedence assumption, that is, on the hypothesis that the constitutive elements of a concept must be an input of the learning process which leads to the acquisition of that concept. In this case my proposal was a model where the constituents of a concept *C* resulted from the same execution of the learning process by virtue of which the concept *C* was acquired. The model consisted in a three-step iterative process, constituted by two general-purpose learning abilities (i.e., dimensional reduction and pattern recognition), and one stage of evaluation and readjustment of the model.

All in all, I think I have presented a plausible explanation of how the acquisition of primitive concepts may happen in a non-circular way. Having said that, this view has left unanswered a range of challenging questions that call for future examination: How plausible is this model from a psychological point of view? Is it possible to provide a characterization of context that articulates the demands of this kind of approach? Or, can be explained the phenomenon of successful communication in a contextualist approach like the one here described? In regard to all this, future research is required, but I am optimistic that forthcoming investigation will give us a deeper understanding of how context may be characterized, and of how successful communication can occur in this type of contextualist framework.

References

Adams, Benjamin, and Martin Raubal. 2009. Conceptual Space Markup Language (CSML): Towards the cognitive semantic web. In *Proceedings of the Third IEEE International Conference on Semantic Computing*. Los Alamitos, CA: IEEE Computer Society Press, 253-260.

Aldenderfer, Mark S., and Roger K. Blashfield. 1984. *Cluster Analysis*. London: Sage Publications.

Allott, Nicholas, and Mark Textor. 2012. Lexical pragmatic adjustment and the nature of ad hoc concepts. *International Review of Pragmatics* 4: 185-208.

Anderson, John R., and Jonathan Betz. 2001. A hybrid model of categorization. *Psychonomic Bulletin & Review* 8: 629-647.

Aquino, Tomás de. 1256-1259. *Quaestiones Disputatae de Veritate*. In *Opera Omnia Iussu Leonis XIII P.M. edita*, tomo 22. Roma: Ad Sanctae Sabinae/Editori di San Tommaso, 1970.

Aristóteles. 1987. *Acerca de la Memoria y de la Reminiscencia*. In *Acerca de la Generación y la Corrupción. Tratados Breves de Historia Natural*. Madrid: Editorial Gredos, 233-255.

Armstrong, Sharon L., Lila R. Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition* 13: 263-308.

Ashby, F. Gregory. 1992. *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Erlbaum Associates.

Ashby, F. Gregory, and James T. Townsend. 1986. Varieties of perceptual independence. *Psychological Review* 93: 154-179.

Ashby, F. Gregory, and Nancy A. Perrin. 1988. Toward a unified theory of similarity and recognition. *Psychological Review* 95: 124-150.

Ashby, F. Gregory, and Ralph E. Gott. 1988. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 33-53.

Ashby, F. Gregory, Leola A. Alfonso-Reese, And U. Turken, and Elliott M. Waldron. 1998. A neuropsychological theory of multiple systems in category learning. *Psychological Review* 105: 442-481.

Attneave, Fred. 1957. Transfer of experience with a class-schema to identification-learning of patterns and shapes. *Journal of Experimental Psychology* 54: 81-88.

Aurenhammer, Franz. 1991. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Computing Surveys* 23: 345-405.

Bailey, Kenneth D. 1975. Cluster analysis. *Sociological Methodology* 6: 59-128.

Barclay, J.R., John D. Bransford, Jeffery J. Franks, Nancy S. McCarrell, and Kathy Nitsch. 1974. Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior* 13: 471-481.

Barlow, Horace B. 1959. Sensory mechanisms, the reduction of redundancy, and intelligence. In D.V. Blake and A.M. Uttley (eds.), *NPL Symposium on the Mechanization of Thought Process* (Vol. 10). London: H.M. Stationery Office, 535-539.

Barsalou, Laurence W. 1983. Ad hoc categories. *Memory & Cognition* 11: 211-227.

—. 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11: 629-654.

—. 1987. The instability of the graded structure: implications for the nature of concepts. In U. Neisser, ed., *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press, 101-140.

—. 1989. Intraconcept similarity and its implications for the interconcept similarity. In S. Vosniadou and A. Ortony, eds., *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 76-121.

—. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In R.S. Wyer Jr. and T.K. Srull, eds., *Content and Process Specificity in the Effects of Prior Experiences. Advances in Social Cognition* (Vol. 3). Hillsdale, NJ: Erlbaum Associates, 61-88.

—. 1992. Frames, concepts, and conceptual fields. In A. Lehrer and E. Feder Kittay, eds., *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale, NJ: Erlbaum Associates, 21-74.

—. 1993. Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A.F. Collins, S.E. Gathercole, M.A. Conway and P.E. Morris, eds., *Theories of Memory*. Hillsdale, NJ: Erlbaum Associates, 29-101.

—. 2012. The human conceptual system. In M.J. Spivey, K. McRae and M.F. Joanisse, eds., *The Cambridge Handbook of Psycholinguistics*. Cambridge: Cambridge University Press, 239-258.

Barsalou, Lawrence W., Janellen Huttenlocher, and Koen Lamberts. 1998. Basing categorization on individuals and events. *Cognitive Psychology* 36: 203-272.

Bartsch, Renate. 1996. The relationship between connectionist models and a dynamic data-oriented theory of concept formation. *Synthese* 108: 421-454.

Bealer, George. 1982. *Quality and Concept*. Oxford: Clarendon Press.

Beals, Richard, David H. Krantz, and Amos Tversky. 1968. Foundations of multidimensional scaling. *Psychological Review* 75: 127-142.

Bechberger, Lucas, and Kai-Uwe Kühnberger. 2017. A thorough formalization of conceptual spaces. In G. Kern-Isberner, J. Fürnkranz and M. Thimm, eds., *KI 2017: Advances in Artificial Intelligence* (Lecture Notes in Computer Science 10505). Cham: Springer, 58-71.

Bechtel, William. 1998. Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science* 22: 295-318.

Bennett, Charles H., Péter Gács, Ming Li, Paul M.B. Vitányi, and Wojciech H. Zurek. 1998. Information distance. *IEEE Transactions on Information Theory* 44: 1407-1423.

Bezdek, James C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Blashfield, Roger K., and Mark S. Aldenderfer. 1988. The methods and problems of cluster analysis. In J.R. Nesselroade and R.B. Cattell, eds., *Handbook of Multivariate Experimental Psychology* (2nd edition). New York: Plenum Press, 447-473.

Bloch-Mullins, Corinne L. 2015. Foundational questions about concepts: Context-sensitivity and embodiment. *Philosophy Compass* 10: 940-952.

—. 2018. Bridging the gap between similarity and causality: An integrated approach to concepts. *British Journal for the Philosophy of Science* 69: 605-632.

Blumenthal, Leonard M. 1953. *Theory and Applications of Distance Geometry*. Oxford: Oxford University Press.

Bock, Hans-Hermann. 2007. Clustering methods: A history of *k*-means algorithms. In P. Brito, P. Bertrand, G. Cucumel and F. de Carvalho, eds., *Selected Contributions in Data Analysis and Classification*. Berlin: Springer, 161-172.

Boom, Jan. 1991. Collective development and the learning paradox. *Human Development* 34: 273-287.

Borg, Ingwer, and Patrick Groenen. 1997. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer.

Bourne, Lyle E, Jr. 1982. Typicality effects in logically defined categories. *Memory & Cognition* 10: 3-9.

Brooks, Lee R. 1978. Nonanalytic concept formation and memory for instances. In E.H. Rosch and B.B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, NJ: Erlbaum Associates, 169-211.

Brooks, Rodney A. 1991. Intelligence without representation. *Artificial Intelligence* 47: 139-159.

Bruner, Jerome S., Jacqueline J. Goodnow, and George A. Austin. 1956. *A Study of Thinking*. Oxford: John Wiley.

Burns, Barbara, and Bryan E. Shepp. 1988. Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception and Psychophysics* 43: 494-507.

Busemann, Herbert. 1955. *The Geometry of Geodesics*. New York: Academic Press.

Bush, Robert R., and Frederick Mosteller. 1951. A model for stimulus generalization and discrimination. *Psychological Review* 58: 413-423.

Cappelen, Herman, and Ernest Lepore. 2005. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

Carey, Susan. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

—. 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

—. 2011. Précis of The Origin of Concepts. *Behavioral and Brain Sciences* 34: 113-167.

—. 2015. Why theories of concepts should not ignore the problem of acquisition. In E. Margolis and S. Laurence, eds., *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press, 415-454.

Carnap, Rudolf. 1923. *Die Quasizerlegung: Ein Verfahren zur Ordnung nichthomogener Mengen mit den Mitteln der Beziehungslehre*. Unpublished manuscript RC-081-04-01, University of Pittsburgh.

—. 1928. *The Logical Structure of the World*. Berkeley, CA: University of California Press, 1967.

—. 1932. Überwindung der Metaphysik durch logische Analyse der Sprache. *Erkenntnis* 2: 219-241.

—. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago, IL: The University of Chicago Press.

Carpenter, Gail A. 1989. Neural network models for pattern recognition and associative memory. *Neural Networks* 2: 243-257.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.

Carroll, Lewis. 1895. What the Tortoise said to Achilles. *Mind* 4: 278-280.

Carruthers, Peter. 1992. *Human Knowledge and Human Nature: A New Introduction to an Ancient Debate*. Oxford: Oxford University Press.

—. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Carston, Robyn. 2002. *Thoughts and Utterances*. London: Blackwell.

Casasanto, Daniel, and Gary Lupyan. 2015. All concepts are ad-hoc concepts. In E. Margolis and S. Laurence, eds., *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press, 543-566.

Catrambone, Richard, and Keith J. Holyoak. 1989. Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15: 1147-1156.

Cattell, Raymond B. 1943. The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology* 38: 476-506.

Cave, Kyle R., Steven Pinker, Liana Giorgi, Catherine E. Thomas, Laurie M. Heller, Jeremy M. Wolfe, and Helen Lin. 1994. The representation of location in visual images. *Cognitive Psychology* 26: 1-32.

Chater, Nick. 1996. Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review* 103: 566-581.

—. 1999. The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology* 52A: 273-302.

Chater, Nick, and Ulrike Hahn. 1997. Representational distortion, similarity and the universal law of generalization. In M. Ramscar, U. Hahn, H. Pain and C. Cambouropoulos, eds., *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization*. Edinburgh: University of Edinburgh. Dept. of Artificial Intelligence, 31-36.

Chater, Nick, and Paul M.B. Vitányi. 2003. The generalized universal law of generalization. *Journal of Mathematical Psychology* 47: 346-369.

Chella, Antonio. 2015. A cognitive architecture for music perception exploiting conceptual spaces. In F. Zenker and P. Gärdenfors, eds., *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Synthese Library* (Vol. 359). Cham: Springer, 187-203.

Chomsky, Noam. 1959. A review of B.F. Skinner's *Verbal Behavior*. *Language* 35: 26-58.

—. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—. 1967. Recent contributions to the theory of innate ideas. *Synthese* 17: 2-11.

—. 1975. *Reflections on Language*. New York: Pantheon Books.

—. 1980. *Rules and Representations*. New York: Columbia University Press.

Churchland, Paul M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

—. 1990. On the nature of theories: A neurocomputational perspective. In K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science* (Vol. 14). Minneapolis: University of Minnesota Press, 59-101.

—. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.

—. 1998. Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy* 95: 5-32.

—. 2005. Chimerical colors: Some phenomenological predictions from cognitive neuroscience. *Philosophical Psychology* 18: 527-560.

Clark, Andy. 1993. *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press.

Clark, Andy, and Josefa Toribio. 1994. Doing without representing? *Synthese* 101: 401-431.

Clark, Austen. 1993. *Sensory Qualities*. Oxford: Clarendon Press.

—. 2000. *A Theory of Sentience*. Oxford: Oxford University Press.

Connolly, Andrew C., Jerry A. Fodor, Lila R. Gleitman, and Henry Gleitman. 2007. Why stereotypes don't even make good defaults. *Cognition* 103: 1-22.

Coombs, Clyde H. 1954. A method for the study of interstimulus similarity. *Psychometrika* 19: 183-194.

Cooper, Gregory F. 1999. An overview of the representation and discovery of causal relationships using Bayesian networks. In C. Glymour and G.F. Cooper, eds., *Computation, Causation, and Discovery*. Cambridge, MA: MIT Press, 3-62.

Costello, Fintan J., and Mark T. Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science* 24: 299-349.

Cowie, Fiona. 1999. *What's Within?: Nativism Reconsidered*. New York: Oxford University Press.

Cox, David R. 1957. Note on grouping. *Journal of the American Statistical Association* 52: 543-547.

Crane, Tim. 1995. *The Mechanical Mind*. London: Routledge.

Cummins, Denise D., and Robert Cummins. 1999. Biological preparedness and evolutionary explanation. *Cognition* 73: B37-B53.

Cummins, Robert. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Czekanowski, Jan. 1911. Objective kriterien in der ethnologie. *Korrespondenz-Blatt der Deutschen Gesselschaft für Anthropologie, Ethnologie, und Urgeschichte* 42: 71-75.

Danks, David. 2014. *Unifying the Mind*. Cambridge, MA: MIT Press.

Davidson, Donald. 1977. Reality without reference. *Dialectica* 31: 247-258.

Decock, Lieven, and Igor Douven. 2011. Similarity after Goodman. *Review of Philosophy and Psychology* 2: 61-75.

Del Pinal, Guillermo. 2016. Prototypes as compositional components of concepts. *Synthese* 193: 2899-2927.

Dennett, Daniel C. 1977. Critical notice: *The Language of Thought* by Jerry Fodor. *Mind* 86: 265-280. (Reprinted as 'A cure for the common code?' in D.C. Dennett, ed., 1979, *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery, VT: Bradford Books, 90-108.)

Descartes, René. 1644. *Principia Philosophiae.* Amsterdam: Elzevir.

—. 1647. *Comments on a Certain Broadsheet*. In J. Cottingham, R. Stoothoff and D. Murdoch, eds., *The Philosophical Writings of Descartes* (Vol. 1). Cambridge: Cambridge University Press, 1985.

Diday, Edwin. 2005. Categorization in symbolic data analysis. In H. Cohen and C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier, 845-867.

Dirichlet, G. Lejeune. 1850. Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *Journal für die Reine und Angewandte Mathematik* 40: 209-227.

Douven, Igor. 2016. Vagueness, graded membership, and conceptual spaces. *Cognition* 151: 80-95.

Douven, Igor, Lieven Decock, Richard Dietz, and Paul Égré. 2013. Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic* 42: 137-160.

Dretske, Fred I. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Driver, Harold E., and Alfred L. Kroeber. 1932. Quantitative expression of cultural relationships. In *University of California Publications in American Archaeology and Ethnology* (Vol. 31). Berkeley: University of California Press, 211-256.

Drobisch, Moritz W. 1855. Uber musikalische Tonbestimmung und Temperatur. *Abhandlungen der Mathematisch-Physischen Classe der Königlich Sächsischen Gesellschaft der Wissenschaften* 4: 1-120.

Dummett, Michael. 1993. *The Seas of Language*. Oxford: Oxford University Press.

Dunn, Joe C. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* 3: 32-57.

Edelman, Shimon. 1995. Representation, similarity, and the chorus of prototypes. *Mind and Machines* 5: 45-68.

—. 1998. Representation is representation of similarities. *Behavioral and Brain Sciences* 21: 449-498.

Edelman, Shimon, and Nathan Intrator. 1997. Learning as extraction of low-dimensional representations. In D.L. Medin, R.L. Goldstone and P.G. Schyns, eds., *The Psychology of Learning and Motivation* (Vol. 36). San Diego: Academic Press, 353-380.

Egan, Frances. 2014. How to think about mental content. *Philosophical Studies* 170: 115-135.

Eisler, Hannes, and Gösta Ekman. 1959. A mechanism of subjective similarity. *Acta Psychologica* 16: 1-10.

Ekman, Gösta, Trygg Engen, Teodor Künnapas, and Ralf Lindman. 1964. A quantitative principle of qualitative similarity. *Journal of Experimental Psychology* 68: 530-536.

Eliasmith, Chris, and Charles H. Anderson. 2003. *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Elman, Jeffrey L. 1990a. Finding structure in time. *Cognitive Science* 14: 179-211.

—. 1990b. Representation and structure in connectionist models. In G.T.M. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press, 345-382.

—. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7: 195-225.

Elman, Jeffrey L., Elisabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Estes, William K. 1986a. Array models for category learning. *Cognitive Psychology* 18: 500-549.

—. 1986b. Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General* 115: 155-174.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Evans, Jonathan St. B.T., and Keith E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8: 223-241.

Evans, Vyvyan. 2006. Lexical concepts, cognitive models and meaning-construction. *Cognitive Linguistics* 17: 491-534.

Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster Analysis*. Chichester: John Wiley.

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4: 272-299.

Fairbanks, Grant, and Patti Grubb. 1961. A psychophysical investigation of vowel formants. *Journal of Speech and Hearing Research* 4: 203-219.

Falkenhainer, Brian, Kenneth D. Forbus, and Dedre Gentner. 1986. The structure-mapping engine. In M. Kaufmann, ed., *Proceedings of the Fifth National Conference on Artificial Intelligence*. Menlo Park: AAAI Press, 272-277.

—. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41: 1-63.

Ferrater Mora, José. 1994. *Diccionario de Filosofía*. Barcelona: Ariel.

Fichet, Bernard, Domenico Piccolo, Rosanna Verde, and Maurizio Vichi. 2011. *Classification and Multivariate Analysis for Complex Data Structures*. Berlin: Springer.

Fillmore, Charles J. 1982. Towards a descriptive framework for spatial deixis. In R.J. Jarvella and W. Klein, eds., *Speech, Place, and Action*. London: John Wiley, 31-59.

Fiorini, S. Rama, Peter Gärdenfors, and Mara Abel. 2014. Representing part-whole relations in conceptual spaces. *Cognitive Processing* 15: 127-142.

Fisher, Walter D. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53: 789-798.

Fodor, Jerry A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

—. 1980a. Fixation of belief and concept acquisition. In M. Piatelli-Palmarini, ed., *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press, 142-149.

—. 1980b. Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3: 63-109.

—. 1981a. The present status of the innateness controversy. In J.A. Fodor, *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 257-316.

—. 1981b. The mind-body problem. *Scientific American* 244: 114-123.

—. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

—. 1990. A theory of content II: The theory. In J.A. Fodor, *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press, 89-136.

—. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

—. 2000. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

—. 2001. Language, thought and compositionality. *Mind& Language* 16: 1-15.

—. 2003. *Hume Variations*. Oxford: Oxford University Press.

—. 2008. *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, Jerry A., Merrill F. Garrett, Edward C.T. Walker, and Cornelia H. Parkes. 1980. Against definitions. *Cognition* 8: 263-367.

Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* 28: 3-71.

Fodor, Jerry A., and Ernest Lepore. 1992. *Holism: A Shopper's Guide*. Cambridge, MA: Blackwell.

—. 1996. The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition* 58: 253-270.

—. 2002. Why meaning (probably) isn't conceptual role. In J.A. Fodor and E. Lepore, *The Compositionality Papers*. Oxford: Oxford University Press, 9-26.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik: Eine Logisch-Mathematische Untersuchung über den Begriff der Zahl*. Breslau: W. Koebner. Translated as *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*, by J.L. Austin, Oxford: Basil Blackwell, 1950.

—. 1892. Uber Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25-50.

—. 1914. Logic in mathematics. In H. Hermes, F. Kambartel and F. Kaulbach, eds., *Posthumous Writings*. Oxford: Basil Blackwell, 203-250.

French, Robert M. 2002. The computational modeling of analogy-making. *Trends in Cognitive Sciences* 6: 200-205.

Galilei, Galileo. 1623. *Il Saggiatore*. Roma: Giacomo Mascardi.

Gärdenfors, Peter. 1990. Induction, conceptual spaces and AI. *Philosophy of Science* 57: 78-95.

—. 1996. Conceptual spaces as a basis for cognitive semantics. In A. Clark, J. Ezquerro and J.M. Larrazabal, eds., *Philosophy and Cognitive Science: Categories, Consciousness, and Reasoning. Philosophical Studies Series* (Vol. 69). Dordrecht: Springer, 159-180.

—. 1999. Some tenets of cognitive semantics. In J. Allwood and P. Gärdenfors, eds., *Cognitive Semantics: Meaning and Cognition. Pragmatics & Beyond New Series* (Vol. 55). Amsterdam: John Benjamins, 19-36.

—. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

—. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press.

Gärdenfors, Peter, and Kenneth Holmqvist. 1994. Concept formation in dimensional spaces. *Lund University Cognitive Studies* 26: 80-95.

Gärdenfors, Peter, and Massimo Warglien. 2012. Using conceptual spaces to model actions and events. *Journal of Semantics* 29: 487-519.

Gärdenfors, Peter, and Frank Zenker. 2013. Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese* 190: 1039-1058.

Garner, Wendell R. 1970. Good patterns have few alternatives. *American Scientist* 58: 34-42.

—. 1974. *The Processing of Information and Structure*. Potomac, MD: Erlbaum Associates.

Garner, Wendell R., and David E. Clement. 1963. Goodness of pattern and pattern uncertainty. *Journal of Verbal Learning and Verbal Behavior* 2: 446-452.

Gauker, Christopher. 2007. A critique of the similarity space theory of concepts. *Mind & Language* 22: 317-345.

—. 2011. *Words and Images: An Essay on the Origin of Ideas*. Oxford: Oxford University Press.

Gauss, Carl F. 1840. Recursion der 'Untersuchungen über die Eigenschaften der positiven ternären quadratischen Formen' von Ludwig August Seeber. *Journal für die Reine und Angewandte Mathematik* 20: 312-320.

Geach, Peter T. 1957. *Mental Acts: Their Content and Their Objects*. London: Routledge & Kegan Paul.

Gelman, Susan A. 2005. *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford: Oxford University Press.

Gelman, Susan A., and Ellen M. Markman. 1986. Categories and induction in young children. *Cognition* 23: 183-209.

Gentner, Dedre. 1980. The structure of analogical models in science. BBN Tech. Report No. 4451. Cambridge, MA: Bolt Beranek and Newman Inc.

—. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7: 155-170.

Gentner, Dedre, and Cecile Toupin. 1986. Systematicity and surface similarity in the development of analogy. *Cognitive Science* 10: 277-300.

Gentner, Dedre, and Arthur B. Markman. 1994. Structural alignment in comparison: No difference without similarity. *Psychological Science* 5: 152-158.

—. 1997. Structure mapping in analogy and similarity. *American Psychologist* 52: 45-56.

Gentner, Dedre, Jeffrey Loewenstein, and Barbara Hung. 2007. Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development* 8: 285-307.

Gettier, Edmund L. 1963. Is justified true belief knowledge? *Analysis* 23: 121-123.

Gick, Mary L., and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15: 1-38.

Gilbert, Charles D., and Wu Li. 2013. Top-down influences on visual processing. *Nature Reviews Neuroscience* 14: 350-363.

Gilmore, Grover C., H. Hersch, Alfonso Camarazza, and J. Griffin. 1979. Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics* 25: 425-431.

Gleitman, Lila R., Andrew C. Connolly, and Sharon L. Armstrong. 2012. Can prototype representations support composition and decomposition? In W. Hinzen, E. Machery and M. Werning, eds., *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, 418-436.

Glock, Hans-Johann. 2010. Concepts, abilities, and propositions. *Grazer Philosophische Studien* 81: 115-134.

Glymour, Clark. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Glymour, Clark, and Gregory F. Cooper. 1999. *Computation, Causation, and Discovery*. Cambridge, MA: MIT Press.

Godfrey-Smith, Peter. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Goldstone, Robert L. 1994a. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 3-28.

—. 1994b. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General* 123: 178-200.

—. 1995. Effects of categorization on color perception. *Psychological Science* 6: 298-304.

—. 1996. Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 988-1001.

Goldstone, Robert L., and Douglas L. Medin. 1994. Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 29-50.

Goldstone, Robert L., Douglas L. Medin, and Jamin Halberstadt. 1997. Similarity in context. *Memory & Cognition* 25: 237-255.

Goldstone, Robert L., and Mark Steyvers. 2001. The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General* 130: 116-139.

Goldstone, Robert L., and Alan Kersten. 2003. Concepts and categorization. In I.B. Weiner, ed., *Handbook of Psychology* (Vol. 4). Hoboken: John Wiley, 599-621.

Goldstone, Robert L., and Ji Yun Son. 2005. Similarity. In K.J. Holyoak and R.G. Morrison, eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 13-36.

Goldstone, Robert L., Joshua R. de Leeuw, and David H. Landy. 2015. Fitting perception in and to cognition. *Cognition* 135: 24-29.

Gonnerman, Chad, and Jonathan M. Weinberg. 2010. Two uneliminated uses for "concepts": Hybrids and guides for inquiry. *Behavioral and Brain Sciences* 33: 211-212.

Goodman, Alvin I. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goodman, Nelson. 1951. *The Structure of Appearance*. Cambridge, MA: Harvard University Press.

—. 1955. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

—. 1967. The epistemological argument. *Synthese* 17: 23-28.

—. 1968. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis: The Bobbs-Merrill Company.

—. 1972. Seven strictures on similarity. In N. Goodman, ed., *Problems and Projects*. Indianapolis: Bobbs-Merrill, 437-447.

Gopnik, Alison, and Henry M. Wellman. 1994. The theory theory. In L.A. Hirschfeld and S.A. Gelman, eds., *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press, 257-293.

Gopnik, Alison, and Andrew N. Meltzoff. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, Alison, and Laura Schulz. 2004. Mechanisms of theory formation in young children. *Trends in Cognitive Sciences* 8: 371-377.

—. 2007. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.

Gordon, A.D. 1987. A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)* 150: 119-137.

—. 1999. *Classification* (2nd edition). Boca Raton: Chapman & Hall/CRC.

Grabmeier, Johannes, and Andreas Rudolph. 2002. Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery* 6: 303-360.

Gruber, Peter M. 2007. *Convex and Discrete Geometry.* Berlin: Springer-Verlag.

Hahn, Ulrike, and Nick Chater. 1997. Concepts and similarity. In K. Lamberts and D. Shanks, eds., *Knowledge, Concepts, and Categories*. East Sussex: Psychology Press, 43-92.

Hahn, Ulrike, Nick Chater, and Lucy B. Richardson. 2003. Similarity as transformation. *Cognition* 87: 1-32.

Hahn, Ulrike, James Close, and Markus Graf. 2009. Transformation direction influences shape-similarity judgments. *Psychological Science* 20: 447-454.

Halford, Graeme S. 2005. Development of thinking. In K.J. Holyoak and R.G. Morrison, eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 529-558.

Hampton, James A. 1979. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior* 18: 441-461.

—. 1987. Inheritance of attributes in natural concept conjunctions. *Memory & Cognition* 15: 55-71.

—. 1991. The combination of prototype concepts. In P.J. Schwanenflugel, ed., *The Psychology of Word Meanings*. Hillsdale, NJ: Erlbaum Associates, 91-116.

—. 1995. Testing the prototype theory of concepts. *Journal of Memory and Language* 34: 686-708.

—. 1997a. Conceptual combination. In K. Lamberts and D. Shanks, eds., *Knowledge, Concepts, and Categories*. Cambridge, MA: MIT Press, 133-159.

—. 1997b. Psychological representation of concepts. In M.A. Conway, ed., *Cognitive Models of Memory*. Cambridge, MA: MIT Press, 81-110.

—. 1998. Similarity-based categorization and fuzziness of natural categories. *Cognition* 65: 137-165.

—. 2006. Concepts as prototypes. In B.H. Ross, ed., *The Psychology of Learning and Motivation* (Vol. 46). New York: Academic Press, 79-113.

—. 2010. Concept talk cannot be avoided. *Behavioral and Brain Sciences* 33: 212-213.

Hampton, James A., and Alessia Passanisi. 2016. When intensions do not map onto extensions: Individual differences in conceptualization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42: 505-523.

Handel, Stephen, and Shiro Imai. 1972. The free classification of analyzable and unanalyzable stimuli. *Perception and Psychophysics* 12: 108-116.

Harman, Harry H. 1967. *Modern Factor Analysis* (2nd edition). Chicago: University of Chicago Press.

Harnad, Stevan. 2005. To cognize is to categorize: Cognition is categorization. In H. Cohen and C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier, 19-43.

Haugeland, John. 1981. Analog and analog. *Philosophical Topics* 12: 213-226.

Hautamäki, Antti. 1992. A conceptual space approach to semantic networks. *Computers & Mathematics with Applications* 23: 517-525.

Heit, Evan, and Joshua Rubinstein. 1994. Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 411-422.

Helmholtz, Hermann von. 1867. *Handbuch der Physiologischen Optik*. Leipzig: Leopold Voss.

Henning, Hans. 1916. *Der Geruch*. Leipzig: Barth.

Hernández-Conde, José V. 2017a. Life cycle of a concept in the ad hoc cognition framework. *Theoria. An International Journal for Theory, History and Foundations of Science* 32: 271-292.

—. 2017b. A case against convexity in conceptual spaces. *Synthese* 194: 4011-4037.

Hinton, Geoffrey E. 1981. Implementing semantic networks in parallel hardware. In G.E. Hinton and J.A. Anderson, eds., *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum Associates, 161-187.

—. 1986. Learning distributed representations of concepts. In *Proceedings of the Eight Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum Associates, 1-12.

—. 1989. Connectionist learning procedures. *Artificial Intelligence* 40: 185-234.

—. 1990. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence* 46: 47-75.

Hinton, Geoffrey E., James L. McClelland, and David E. Rumelhart. 1986. Distributed representations. In D.E. Rumelhart and J.L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* (Vol. 1). Cambridge, MA: MIT Press, 77-109.

Hinton, Geoffrey E., Peter Dayan, Brendan J. Frey, and Radford M. Neal. 1995. The "wake-sleep" algorithm for unsupervised neural networks. *Science* (New Series) 268: 1158-1161.

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313: 504-507.

Hintzman, Douglas L. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychological Review* 93: 411-428.

Hintzman, Douglas L., and Genevieve Ludlam. 1980. Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition* 8: 378-382.

Hodgetts, Carl J. 2011. Transformation and representation in similarity. Ph.D. dis., Cardiff University.

Hodgetts, Carl J., Ulrike Hahn, and Nick Chater. 2009. Transformation and alignment in similarity. *Cognition* 113: 62-79.

Hoenig, Klaus, Eun-Jin Sim, Viktor Bochey, Bärbel Herrnberger, and Markus Kiefer. 2008. Conceptual flexibility in the human brain: Dynamic recruitment of semantic maps from visual, motor, and motion-related areas. *Journal of Cognitive Neuroscience* 20: 1799-1814.

Hofstadter, Douglas R. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Hofstadter, Douglas R., and Melanie Mitchell. 1994. The Copycat project: A model of mental fluidity and analogy-making. In K.J. Holyoak and J.A. Barnden, eds. *Analogical Connections: 002 (Advances in Connectionist and Neural Computation Theory*. Norwood, NJ: Ablex, 31-112.

Holman, Eric W. 1979. Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology* 20: 1-15.

Holyoak, Keith J. 2005. Analogy. In K.J. Holyoak and R.G. Morrison, eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 117-142.

Holyoak, Keith J., and Peter C. Gordon. 1983. Social reference points. *Journal of Personality and Social Psychology* 44: 881-887.

Holyoak, Keith J., and Paul Thagard. 1989. Analogical mapping by constraint satisfaction. *Cognitive Science* 13: 295-355.

—. 1995. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Homa, Donald, Sharon Sterling, and Lawrence Trepel. 1981. Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory* 104: 418-439.

Homa, Donald, Sherry Dunbar, and Liva Nohre. 1991. Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17: 444-458.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417-441, 498-520.

Hull, Clark L. 1928. Quantitative aspects of the evolution of concepts. *Psychological Monographs* 28: 1-86.

Hume, David. 1739. *A Treatise of Human Nature*. L.A. Selby-Bigge and P.H. Nidditch, eds. Oxford: Oxford University Press, 1978.

—. 1741. *An Enquiry Concerning Human Understanding*. P. Millican, ed. Oxford: Oxford University Press, 2007.

Hummel, John E., and Keith J. Holyoak. 1997. Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 3: 427-466.

Hunt, Earl B. 1962. *Concept Learning: An Information Processing Approach*. New York: John Wiley.

Hyman, Ray, and Arnold Well. 1968. Perceptual separability and spatial models. *Perception and Psychophysics* 3: 161-165.

Imai, Shiro. 1977. Pattern similarity and cognitive transformations. *Acta Psychologica* 41: 433-447.

Jackendoff, Ray S. 1983. *Semantics and Cognition*. Cambridge, MA: MIT Press.

—. 1989. What is a concept, that a person may grasp it? *Mind & Language* 4: 68-102.

Jackson, J. Edward. 1991. *A User's Guide to Principal Components*. New York: John Wiley.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31: 651-666.

Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall.

Jain, Anil K., Robert P.W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22: 4-37.

James, William. 1890. *The Principles of Psychology*. New York: Henry Holt.

Jensen, Arthur R. 1998. *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Johannesson, Mikael. 2000. Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology* 53: 121-139.

Jolliffe, Ian T. 2002. *Principal Component Analysis* (2nd edition). New York: Springer.

Jönsson, Martin L., and James A. Hampton. 2007. On prototypes as defaults (Comment on Connolly, Fodor, Gleitman and Gleitman, 2007). *Cognition* 106: 913-923.

Kamp, Hans, and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition* 57: 129-191.

Katz, Jerrold J., and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39: 170-210.

Keane, Mark T., Tim Ledgeway, and Stuart Duff. 1994. Constraints on analogical mapping: A comparison of three models. *Cognitive Science* 18: 387-438.

Keil, Frank C. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

—. 1994. The birth and nurturance of concepts by domains: The origins of concepts of living things. In L.A. Hirschfeld and S.A. Gelman, eds., *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press, 234-254.

Keil, Frank C., W. Carter Smith, Daniel J. Simons, and Daniel T. Levin. 1998. Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition* 65: 103-135.

Kelley, John L. 1955. *General Topology*. New York: Van Nostrand.

Kemler, Deborah G., and Linda B. Smith. 1978. Is there a developmental trend from integrality to separability in perception? *Journal of Experimental Child Psychology* 26: 498-507.

Kenny, Anthony. 2010. Concepts, brains, and behaviour. *Grazer Philosophische Studien* 81: 105-113.

Kiefer, Markus, and Friedemann Pulvermüller. 2012. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex* 48: 805-825.

Kintsch, Walter. 1974. *The Representation of Meaning in Memory*. Hillsdale, NJ: Erlbaum Associates.

Knapp, Andrew G., and James A. Anderson. 1984. Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19: 616-637.

Kohonen, Teuvo. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59-69.

Kolmogorov, Andréi. 1963. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A* 25: 369-376.

—. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission* 1: 1-7.

Komatsu, Lloyd K. 1992. Recent views of conceptual structure. *Psychological Bulletin* 112: 500-526.

Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

—. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Krumhansl, Carol L. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review* 85: 445-463.

Kruschke, John K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99: 22-44.

—. 2005. Category learning. In K. Lamberts and R. Goldstone, eds., *Handbook of Cognition*. London: Sage Publications, 183-201.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: The University of Chicago Press.

Lalumera, Elisabetta. 2010. Concepts are a functional kind. *Behavioral and Brain Sciences* 33: 217-218.

Lamberts, Koen. 1998. The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24: 695-711.

—. 2000. Information-accumulation theory of speeded classification. *Psychological Review* 107: 227-260.

Landau, Barbara. 1982. Will the real grandmother please stand up? The psychological reality of dual meaning representations. *Journal of Psycholinguistic Research* 11: 47-62.

Laurence, Stephen, and Eric Margolis. 1999. Concepts and cognitive science. In E. Margolis and S. Laurence, eds., *Concepts: Core Readings*. Cambridge, MA: MIT Press, 3-81.

—. 2001. The poverty of the stimulus argument. *British Journal for the Philosophy of Science* 52: 217-276.

—. 2002. Radical concept nativism. *Cognition* 86: 25-55.

—. 2015. Concept nativism and neural plasticity. In E. Margolis and S. Laurence, eds., *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press, 117-147.

Lawley, Derrick N., and Albert E. Maxwell. 1971. *Factor Analysis as a Statistical Method* (2nd edition). London: Butterworth.

Lebois, Lauren A.M., Christine D. Wilson-Mendenhall, and Lawrence W. Barsalou. 2015. Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science* 39: 1764-1801.

Lee, Daniel D., and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788-791.

—. 2001. Algorithms for non-negative matrix factorization. In T.K. Leen, T.G. Dietterich and V. Tresp, eds., *Proceedings of the 13th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press, 535-541.

Lee, Der-Tsai. 1980. Two-dimensional Voronoi diagrams in the L_p-metric. *Journal of the Association for Computing Machinery* 27: 604-618.

Leibniz, Gottfried W. 1765. *Nouveaux Essais Sur L'entendement Humain*. In R.E. Raspe, ed., *Œuvres Philosophiques latines & françoises de feu M. de Leibnitz*. Amsterdam and Leipzig: J. Schreuder.

Leibniz, Gottfried W. (A). *Sämtliche Schriften und Briefe* (Series 1-7). Berlin-Brandenburgischen Akademie der Wissenschaften and Akademie der Wissenschaften in Göttingen, ed. Berlin: Akademie Verlag, 1923. Cited by series, volume, and page (e.g., "A64 107" refers to series 6, volume 4, page 107).

Leitgeb, Hannes. 2005. How similarities compose. In M. Werning, E. Machery and G. Schurz, eds., *The Compositionality of Meaning and Content I: Foundational Issues*. Frankfurt: Ontos Press, 147-168.

—. 2007. A new analysis of quasianalysis. *Journal of Philosophical Logic* 36: 181-226.

Lewis, David. 1971. Analog and digital. *Noûs* 5: 321-327.

Li, Ming, and Paul Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer.

Lloyd, Stuart P. 1957. Least squares quantization in PCM. Unpublished *Bell Laboratories Technical Note*. Partly presented at the Institute of Mathematical Statistics Meeting, Atlantic City, NJ (September 1957). Published in 1982, *IEEE Transaction on Information Theory* 28: 129-137.

Locke, John. 1690. *An Essay Concerning Human Understanding*. P.H. Nidditch, ed. Oxford: Oxford University Press, 1975.

Loewenstein, Jeffrey, and Dedre Gentner. 2001. Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development* 1: 189-219.

—. 2005. Relational language and the development of relational mapping. *Cognitive Psychology* 50: 315-353.

Löhr, Guido. 2017. Abstract concepts, compositionality, and the contextualism-invariantism debate. *Philosophical Psychology* 30: 689-710.

Love, Bradley C. 2000. A computational level theory of similarity. In L.R. Gleitman and A.K. Joshi, eds., *Proceedings of the Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum Associates, 316-321.

Love, Bradley C., Douglas L. Medin, and Todd M. Gureckis. 2004. SUSTAIN: A network model of category learning. *Psychological Review* 111: 309-332.

Machery, Edouard. 2005. Concepts are not a natural kind. *Philosophy of Science* 72: 444-467.

—. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

—. 2010a. The heterogeneity of knowledge representation and the elimination of *concept*. *Behavioral and Brain Sciences* 33: 231-239.

—. 2010b. Reply to Barbara Malt and Jesse Prinz. *Mind & Language* 25: 634-646.

—. 2015. By default: Concepts are accessed in a context-independent manner. In E. Margolis and S. Laurence, eds., *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press, 567-588.

Machery, Edouard, and Selja Seppälä. 2011. Against hybrid theories of concepts. *Anthropology & Philosophy* 1: 99-127.

Machery, Edouard, and Lisa G. Lederer. 2012. Simple heuristics for concept composition. In W. Hinzen, E. Machery and M. Werning, eds., *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, 454-472.

MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. In L.M. Le Cam and J. Neyman, eds., *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley: University of California Press, 281-297.

Maddox, W. Todd. 1992. Perceptual and decisional separability. In G.F. Ashby, ed., *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Erlbaum Associates, 147-180.

Maddox, W. Todd, and F. Gregory Ashby. 1993. Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics* 53: 49-70.

Maley, Corey J. 2011. Analog and digital, continuous and discrete. *Philosophical Studies* 155: 117-131.

Malt, Barbara C. 1994. Water is not H_2O. *Cognitive Psychology* 27: 41-70.

—. 2010. Why we should do without concepts. *Mind & Language* 25: 622-633.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Marcus, Gary F. 1998. Can connectionism save constructivism? *Cognition* 66: 153-182.

Margolis, Eric. 1994. A reassessment of the shift from the classical theory of concepts to prototype theory. *Cognition* 51: 73-89.

—. 1998. How to acquire a concept. *Mind & Language* 13: 347-369.

Margolis, Eric, and Stephen Laurence. 2003. Concepts. In S.P. Stich and T.A. Warfield, eds., *The Blackwell Guide to Philosophy of Mind*. Oxford: Blackwell, 190-213.

—. 2007. The ontology of concepts: Abstract objects or mental representations? *Noûs* 41: 561-593.

—. 2010. Concepts and theoretical unification. *Behavioral and Brain Sciences* 33: 219-220.

—. 2011a. Concepts. In E.N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). https://plato.stanford.edu/archives/spr2014/entries/concepts/.

—. 2011b. Beyond the building blocks model. *Behavioral and Brain Sciences* 34: 139-140.

—. 2011c. Learning matters: The role of learning in concept acquisition. *Mind & Language* 26: 507-539.

—. 2013. In defense of nativism. *Philosophical Studies* 165: 693-718.

Markman, Arthur B., and Dedre Gentner. 1993a. Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language* 32: 517-535.

—. 1993b. Structural alignment during similarity comparisons. *Cognitive Psychology* 25: 431-467.

Marr, David. 1982. *Vision*. San Francisco: W.H. Freeman and Company.

Marr, David, and H. Keith Nishihara. 1978. Representation and recognition of spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London, Series B, Biological Sciences* 200: 269-294.

Matthen, Mohan. 2005. *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*. Oxford: Clarendon Press.

Mazzone, Marco, and Elisabetta Lalumera. 2010. Concepts: Stored or created? *Minds & Machines* 20: 47-68.

McClelland, James L., and David E. Rumelhart. 1985. Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* 114: 159-188.

McClelland, James L., Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Science* 14: 348-356.

McCulloch, Warren S., and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5: 115-133.

Medin, Douglas L. 1986. Comment on "Memory storage and retrieval processes in category learning". *Journal of Experimental Psychology: General* 115: 373-381.

—. 1989. Concepts and conceptual structure. *American Psychologist* 44: 1469-1481.

Medin, Douglas L., and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85: 207-238.

Medin, Douglas L., and Edward J. Shoben. 1988. Context and structure in conceptual combination. *Cognitive Psychology* 20: 158-190.

Medin, Douglas L., Robert L. Goldstone, and Dedre Gentner. 1993. Respects for similarity. *Psychological Review* 100: 254-278.

Melara, Robert D. 1992. The concept of perceptual similarity: From psychophysics to cognitive psychology. In D. Algom, ed., *Psychophysical Approaches to Cognition*. Amsterdam: Elsevier, 303-388.

Melara, Robert D., Lawrence E. Marks, and Kathryn E. Lesko. 1992. Optional processes in similarity judgments. *Perception and Psychophysics* 51: 123-133.

Mervis, Carolyn B., and Eleanor H. Rosch. 1981. Categorization of natural objects. *Annual Review of Psychology* 32: 89-115.

Miller, George A., and Philip N. Johnson-Laird. 1976. *Language and Perception*. Cambridge, MA: Harvard University Press.

Millikan, Ruth G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

—. 1998. A common structure for concepts of individuals, stuffs, and real kinds: More Mama, more milk, and more mouse. *Behavioral and Brain Sciences* 21: 55-100.

—. 2000. *On Clear and Confused Ideas: An Essay about Substance Concepts*. Cambridge: Cambridge University Press.

Mirkin, Boris G. 1987. Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification* 4: 7-31.

Mitchell, Melanie. 1993. *Analogy-Making as Perception: A Computer Model*. Cambridge, MA: MIT Press.

Molenaar, Peter C.M. 1986. On the impossibility of acquiring more powerful structures: A neglected alternative. *Human Development* 29: 245-251.

Morgan, Michael J. 2016. Feature analysis. In M.A. Arbib, ed., *The Handbook of Brain Theory and Neural Networks* (2nd edition). Cambridge, MA: MIT Press, 444-449.

Mormann, Thomas. 1994. A representational reconstruction of Carnap's quasianalysis. In R.M. Burian, D. Hull and M. Forbes, eds., *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Chicago, IL: The University of Chicago Press, 96-104.

—. 2009. New work for Carnap's quasi-analysis. *Journal of Philosophical Logic* 38: 249-282.

Mozer, Michael C. 1994. Computational approaches to functional feature learning. In A. Ram and K. Eiselt, eds., *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 975-976.

Murphy, Gregory L. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.

—. 2016. Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review* 23: 1035-1042.

Murphy, Gregory L., and Douglas L. Medin. 1985. The role of theories in conceptual coherence. *Psychological Review* 92: 289-316.

Napoli, Amedeo. 2005. A smooth introduction to symbolic methods for knowledge discovery. In H. Cohen and C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier, 913-933.

Neimark, Edith D. 1983. There is one classification system with a long developmental history. In E.K. Scholnick, ed., *New Trends in Conceptual Representation: Challenges to Piaget's Theory*. Hillsdale, NJ: Erlbaum Associates, 111-127.

Newell, Allen. 1980. Physical symbol systems. *Cognitive Science* 4: 135-183.

Newell, Allen, and Herbert A. Simon. 1976. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM* 19: 113-126.

Newton, Isaac. 1704. *Opticks*. London: Smith and Walford.

Niiniluoto, Ilkka. 1987. *Truthlikeness*. Dordrecht: Reidel Publishing Company.

Nosofsky, Robert M. 1984. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10: 104-114.

—. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115: 39-57.

—. 1987. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13: 87-108.

—. 1988a. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 700-708.

—. 1988b. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 54-65.

—. 1991. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology* 23: 94-140.

—. 1992a. Exemplar-based approach to relating categorization, identification, and recognition. In F.G. Ashby, ed., *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Erlbaum Associates, 363-393.

—. 1992b. Similarity scaling and cognitive process models. *Annual Review of Psychology* 43: 25-53.

—. 2011. The generalized context model: An exemplar model of classification. In E.M. Pothos and A.J. Wills, eds., *Formal Approaches in Categorization*. Cambridge: Cambridge University Press, 18-39.

Nosofsky, Robert M., Thomas J. Palmeri, and Stephen C. McKinley. 1994. Rule-plus-exception model of classification learning. *Psychological Review* 101: 53-79.

Nosofsky, Robert M., and Thomas J. Palmeri. 1997. An exemplar-based random walk model of speeded classification. *Psychological Review* 104: 266-300.

Nosofsky, Robert M., and Safa R. Zaki. 2002. Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28: 924-940.

Oden, Gregg C. 1977. Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance* 3: 565-575.

Ogden, Charles K., and Ivor A. Richards. 1923. *The Meaning of Meaning*. New York: Harcourt, Brace & World, Inc.

Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Sung N. Chiu. 1992. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. New York: John Wiley.

Osherson, Daniel N., and Edward E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9: 35-58.

Osherson, Daniel N., Edward E. Smith, Ormond Wilkie, Alejandro López, and Eldar Shafir. 1990. Category-based induction. *Psychological Review* 97: 185-200.

Pagin, Peter. 2012. Communication and the complexity of semantics. In M. Werning, W. Hinzen and E. Machery, eds., *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, 510-529.

Palmer, Stephen E. 1983. The psychology of perceptual organization: A transformational approach. In J. Beck, B. Hope and A. Rosenfeld, eds., *Human and Machine Vision*. New York: Academic Press, 269-339.

Palmeri, Thomas J. 1997. Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23: 324-354.

Paradis, Carita. 2015. Conceptual spaces at work in sensory cognition: Domains, dimensions and distances. In F. Zenker and P. Gärdenfors, eds., *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Synthese Library* (Vol. 359). Cham: Springer, 33-55.

Peacocke, Christopher. 1992. *A Study of Concept*. Cambridge, MA: MIT Press.

Pearson, Karl. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2: 559-572.

Piaget, Jean. 1937. *La Construction du Réel chez l'Enfant*. Neuchâtel: Delachaux et Niestlé. Translated in M. Cook (trans.), *The Construction of Reality in the Child*. New York: Basic Books, 1954.

Piccinini, Gualtiero, and Sam Scott. 2006. Splitting concepts. *Philosophy of Science* 73: 390-409.

Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

—. 2002. *The Blank Slate: The Modern Denial of Human Nature*. New York: Viking.

—. 2007. *The Stuff of Thought: Language as a Window into Human Nature*. New York: Viking.

Pinker, Steven, and Alan Prince. 1996. The nature of human concepts: Evidence from an unusual source. *Communication and Cognition* 29: 307-362.

Pitt, David. 2017. Mental representation. In E.N. Zalta ed., *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), https://plato.stanford.edu/archives/spr2017/entries/mental-representation/.

Plato. *Parmenides*. In J. Burnet, ed., *Platonis Opera* (Vol. 2). Oxford: Clarendon Press, 1901.

Podgorny, Peter, and Wendell R. Garner. 1979. Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics* 26: 37-52.

Posner, Michael I. 1964. Information reduction in the analysis of sequential tasks. *Psychological Review* 71: 491-504.

—. 1969. Abstraction and the process of recognition. In G.H. Bower and J.T. Spence, eds., *The Psychology of Learning and Motivation* (Vol. 3). New York: Academic Press, 44-100.

Posner, Michael I., and Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77: 353-363.

Price, Henry H. 1953. *Thinking and Experience*. London: Hutchinson.

Prinz, Jesse J. 2002. *Furnishing the Mind*. Cambridge, MA: MIT Press.

—. 2010. Can concept empiricism forestall eliminativism? *Mind & Language* 25: 612-621.

—. 2012. Regaining composure: A defense of prototype compositionality. In W. Hinzen, E. Machery and M. Werning, eds., *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, 437-453.

Putnam, Hilary. 1967. The 'innateness hypothesis' and explanatory models in linguistics. *Synthese* 17: 12-22.

—. 1970. Is semantics possible? *Metaphilosophy* 1: 187-201.

—. 1975. The meaning of 'meaning'. In K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science* (Vol. 7). Minneapolis: University of Minnesota Press, 131-193.

—. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.

Pylyshyn, Zenon W. 1980. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3: 111-169.

Quine, Willard V. 1951. Two dogmas of empiricism. *The Philosophical Review* 60: 20-43.

—. 1969. Natural kinds. In *Ontological Relativity and Other Essays*. New York: Columbia University Press, 114-138.

Quinlan, J. Ross, and Ronald L. Rivest. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80: 227-248.

Ramsey, William M. 1997. Do connectionist representations earn their explanatory keep? *Mind & Language* 12: 34-66.

—. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press.

—. 2017. Must cognition be representational? *Synthese* 194: 4197-4214.

Ramsey, William M., Stephen P. Stich, and Joseph Garon. 1990. Connectionism, eliminativism, and the future of folk psychology. *Philosophical Perspectives* 4: 499-533.

Recanati, François. 2002. The Fodorian fallacy. *Analysis* 62: 285-289.

—. 2012. *Mental Files*. Oxford: Oxford University Press.

Redlich, A. Norman. 1993. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation* 5: 289-304.

Reed, Stephen K. 1972. Pattern recognition and categorization. *Cognitive Psychology* 3: 382-407.

Rehder, Bob. 2003a. Categorization as causal reasoning. *Cognitive Science* 27: 709-748.

—. 2003b. A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 1141-1159.

Rehder, Bob, and Reid Hastie. 2001. Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General* 130: 323-360.

Rey, Georges. 1983. Concepts and stereotypes. *Cognition* 15: 237-262.

—. 1994. Concepts. In S. Guttenplan, ed., *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 185-193.

—. 2014. Innate and learned: Carey, mad dog nativism, and the poverty of stimuli and analogies (yet again). *Mind & language* 29: 109-132.

Rice, Collin. 2016. Concepts as pluralistic hybrids. *Philosophy and Phenomenological Research* 92: 597-619.

Rips, Lance J. 1989. Similarity, typicality, and categorization. In S. Vosniadou and A. Ortony, eds., *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 21-59.

Rips, Lance J., Amber Bloomfield, and Jennifer Asmuth. 2008. From numerical concepts to concepts of number. *Behavioral and Brain Sciences* 31: 623-642.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14: 465-471.

—. 1989. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Ritter, Helge, and Teuvo Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61: 241-254.

Rogers, Timothy T., and James L. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Rolls, Edmund T. 2008. Top-down control of visual perception: Attention in natural vision. *Perception* 37: 333-354.

Rosch, Eleanor H. 1973. On the internal structure of perceptual and semantic categories. In T. Moore, ed., *Cognitive Development and the Acquisition of Learning*. New York: Academic Press, 111-144.

—. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104: 192-233.

—. 1978. Principles of categorization. In E.H. Rosch and B.B. Lloyd, eds., *Cognition and Categorization*. Hillsdale: Erlbaum Associates, 27-48.

—. 1983. Prototype classification and logical classification: The two systems. In E.K. Scholnick, ed., *New Trends in Conceptual Representation: Challenges to Piaget's Theory*. Hillsdale, NJ: Erlbaum Associates, 73-86.

Rosch, Eleanor H., and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7: 573-605.

Rosch, Eleanor H., Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8: 382-439.

Ross, Brian H. 1987. This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13: 629-639.

Roth, Emilie M., and Edward J. Shoben. 1983. The effect of context on the structure of categories. *Cognitive Psychology* 15: 346-378.

Rowlands, Mark. 2017. Arguing about representation. *Synthese* 194: 4215-4232.

Rumelhart, David E. 1989. The architecture of mind: A connectionist approach. In M.I. Posner, ed., *Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 133-159.

Rumelhart, David E., and James L. McClelland. 1986. *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Ryder, Dan. 2009. Problems of representation I: Nature and role. In J. Symons and P. Calvo, eds., *The Routledge Companion to Philosophy of Psychology*. Abingdon: Routledge, 233-250.

Samet, Jerry, and Owen Flanagan. 1989. Innate representations. In S. Silvers, ed., *Representation: Readings in the Philosophy of Mental Representation. Philosophical Studies Series* (Vol. 40). Dordrecht: Springer, 189-210.

Samuels, Richard, and Michael Ferreira. 2010. Why *don't* concepts constitute a natural kind? *Behavioral and Brain Sciences* 33: 222-223.

Schreider, Julius A. 1975. *Equality, Resemblance and Order*. Moscow: Mir Publishers.

Schrödinger, Erwin. 1920. Grundlinien einer Theorie der Farbenmetrik im Tagessehen der Farbenmetrik II. Teil: Höhere Farbenmetrik (eigentliche Metrik der Farbe). *Annalen der Physik* 63: 481-520.

Schurz, Gerhard. 2012. Prototypes and their compositionality from an evolutionary point of view. In W. Hinzen, E. Machery and M. Werning, eds., *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, 530-553.

Schyns, Philippe G. 1991. A modular neural network model of concept acquisition. *Cognitive Science* 15: 461-508.

Schyns, Philippe G., and Luc Rodet. 1997. Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23: 681-696.

Schyns, Philippe G., Robert L. Goldstone, and Jean-Pierre Thibaut. 1998. The development of features in object concepts. *Behavioral and Brain Sciences* 21: 1-54.

Searle, John R. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT 'Press.

Seligman, Martin E.P. 1971. Phobias and preparedness. *Behavior Therapy* 2: 307-320.

Shagrir, Oron. 2012. Structural representations and the brain. *British Journal for the Philosophy of Science* 63: 519-545.

Shepard, Roger N. 1957. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika* 22: 325-345.

—. 1958. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology* 55: 509-523.

—. 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* 27: 125-140.

—. 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* 27: 219-246.

—. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* (New Series) 210: 390-398.

—. 1987. Toward a universal law of generalization for psychological science. *Science* 237: 1317-1323.

Shepard, Roger N., and Phipps Arabie. 1979. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* 86: 87-123.

Sivik, Lars, and Charles Taft. 1994. Color naming: A mapping in the NCS of common color terms. *Scandinavian Journal of Psychology* 35: 144-164.

Sjöberg, Lennart. 1972. A cognitive theory of similarity. *Göteborg Psychological Reports* 2(10).

Sjöberg, Lennart, and Christer Thorslund. 1979. A classificatory theory of similarity. *Psychological Research* 40: 223-247.

Sloman, Steven A. 1993. Feature-based induction. *Cognitive Psychology* 25: 231-280.

—. 1996. The empirical case for two systems of reasoning. *Psychological Review* 119: 3-22.

—. 1998. Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology* 35: 1-33.

—. 2005. *Causal Models: How People Think about the World and Its Alternatives*. Oxford: Oxford University Press.

Smith, Edward E. 1989. Concepts and induction. In M.I. Posner, ed., *Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 501-526.

Smith, Edward E., Edward J. Shoben, and Lance J. Rips. 1974. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review* 81: 214-241.

Smith, Edward E., and Douglas L. Medin. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Smith, Edward E., Douglas L. Medin, and Lance J. Rips. 1984. A psychological approach to concepts: Comments on Rey's "Concepts and stereotypes". *Cognition* 17: 265-274.

Smith, Edward E., and Daniel N. Osherson. 1984. Conceptual combination with prototype concepts. *Cognitive Science* 8: 337-361.

—. 1989. Similarity and decision making. In S. Vosniadou and A. Ortony, eds., *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 60-75.

Smith, Edward E., Daniel N. Osherson, Lance J. Rips, and Margaret Keane. 1988. Combining prototypes: A selective modification model. *Cognitive Science* 12: 485-527.

Smith, Edward E., and Steven A. Sloman. 1994. Similarity- versus rule-based categorization. *Memory & Cognition* 22: 377-386.

Smith, Edward E., Andrea L. Patalano, and John Jonides. 1998. Alternative strategies of categorization. *Cognition* 65: 167-196.

Smith, J. David, and Deborah G. Kemler. 1984. Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General* 113: 137-159.

Smith, J. David, and John P. Minda. 2002. Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28: 800-811.

Smith, Linda B. 1989. From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou and A. Ortony, eds., *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 146-178.

Smith, Linda B., and Deborah G. Kemler. 1978. Levels of experienced dimensionality in children and adults. *Cognitive Psychology* 10: 502-532.

Smolensky, Paul. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11: 1-74.

—. 1991. Connectionism, constituency, and the language of thought. In B. Lower and G. Rey, eds., *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell, 201-227.

Sokal, Robert R., and Peter H.A. Sneath. 1963. *Principles of Numerical Taxonomy*. San Francisco: W.H. Freeman and Co.

Solomon, Karen O., Douglas L. Medin, and Elisabeth Lynch. 1999. Concepts do more than categorize. *Trends in Cognitive Sciences* 3: 99-105.

Spearman, Charles. 1904. "General intelligence," objectively determined and measured. *American Journal of Psychology* 15: 201-292.

—. 1927. *The Abilities of Man: Their Nature and Measurement*. London: Macmillan.

Spelke, Elisabeth S. 1998. Nativism, empiricism, and the origins of thought. *Infant Behavior & Development* 21: 181-200.

Spelke, Elisabeth S., and Katherine D. Kinzler . 2007. Core knowledge. *Developmental Science* 10: 89-96.

Sperber, Dan, and Deirdre Wilson. 1995. *Relevance: Communication and Cognition* (2nd edition). Oxford: Blackwell.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Steinhaus, Hugo. 1956. Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences. Class III* 4: 801-804.

Sternberg, Robert J. 2005. Intelligence. In K.J. Holyoak and R.G. Morrison, eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 751-773.

Stewart, G.W. 1993. On the early history of the singular value decomposition. *SIAM Review* 4: 551-566.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief.* Cambridge, MA: MIT Press.

—. 1992. What is a theory of mental representation? *Mind* 101: 243-261.

Strohminger, Nina, and Bradley W. Moore. 2010. Banishing the thought. *Behavioral and Brain Sciences* 33: 225-226.

Sutton, Jonathan. 2004. Are concepts mental representations or abstracta? *Philosophy and Phenomenological Research* 68: 89-108.

Takane, Yoshio, and Justine Sergent. 1983. Multidimensional scaling models for reaction times and same-different judgments. *Psychometrika* 48: 393-423.

Tarr, Michael J. 1995. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review* 2: 55-82.

Tarr, Michael J., and Steven Pinker. 1989. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology* 21: 233-282.

Tenenbaum, Joshua B. 1999. Bayesian modeling of human concept learning. In M.S. Kearns, S.A. Solla and D.A. Cohn, eds., *Advances in Neural Information Processing Systems 11*. Cambridge, MA: MIT Press, 59-68.

Tenenbaum, Joshua B., Vin de Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290: 2319-2323.

Tenenbaum, Joshua B., and Thomas L. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24: 629-640.

Thomas, Michael S.C., and Denis Mareschal. 2001. Metaphor as categorization: A connectionist implementation. *Metaphor and Symbol* 16: 5-27.

Thomas, Michael S.C., Harry R.M. Purser, and Denis Mareschal. 2012. Is the mystery of thought demystified by context-dependent categorisation? Towards a new relation between language and thought. *Mind & Language* 27: 595-618.

Thorpe, Simon J. 2003. Localized versus distributed representations. In M.A. Arbib, ed., *The Handbook of Brain Theory and Neural Networks* (2nd edition). Cambridge, MA: MIT Press, 643-646.

Thurstone, Louis Leon. 1931. Multiple factor analysis. *Psychological Review* 38: 406-427.

—. 1947. *Multiple Factor Analysis: A Development and Expansion of* The Vectors of Mind. Chicago, IL: The University of Chicago Press.

Torgerson, Warren S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17: 401-419.

—. 1958. *Theory and Methods of Scaling*. New York: John Wiley.

—. 1965. Multidimensional scaling of similarity. *Psychometrika* 30: 379-393.

Tryon, Robert C. 1939. *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor, MI: Edwards Brothers.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59: 433-460.

Turk, Matthew, and Alex Pentland. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3: 71-86.

Tversky, Amos. 1977. Features of similarity. *Psychological Review* 84: 327-352.

Tversky, Amos, and David H. Krantz. 1970. The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology* 7: 572-596.

Tversky, Amos, and Imatar Gati. 1978. Studies of similarity. In E.H. Rosch and B.B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, NJ: Erlbaum Associates, 79-98.

—. 1982. Similarity, separability, and the triangle inequality. *Psychological Review* 89: 123-154.

Ullman, Shimon. 1996. *High-level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press.

Van Gelder, Tim. 1995. What might cognition be, if not computation? *Journal of Philosophy* 92: 345-381.

Vanpaemel, Wolf, and Gert Storms. 2010. Abstraction and model evaluation in category learning. *Behavior Research Methods* 42: 421-437.

Vicente, Agustín, and Fernando Martínez-Manrique. 2016. The big concepts paper: A defence of hybridism. *British Journal for the Philosophy of Science* 67: 59-88.

Vitányi, Paul M.B., Frank J. Balbach, Rudi L. Cilibrasi, and Ming Li. 2009. Normalized information distance. In F. Emmert-Streib and M. Dehmer, eds. *Information Theory and Statistical Learning*. New York: Springer, 45-82.

Von der Malsburg, Christoph. 1973. Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetik* 14: 85-100.

Voronoi, Georges. 1908. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik* 134: 198-287.

Yee, Eiling, and Sharon L. Thompson-Schill. 2016. Putting concepts into context. *Psychonomic Bulletin & Review* 23: 1015-1027.

Wallace, Christopher S., and David M. Boulton. 1968. An information measure for classification. *The Computer Journal* 11: 185-194.

Wallace, Christopher S., and David L. Dowe. 1999. Minimum message length and Kolmogorov complexity. *The Computer Journal* 42: 270-283.

Ward, Thomas B. 1983. Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance* 9: 103-112.

Warglien, Massimo, and Peter Gärdenfors. 2013. Semantics, conceptual spaces, and the meeting of minds. *Synthese* 190: 2165-2193.

—. 2015. Meaning negotiation. In F. Zenker and P. Gärdenfors, eds., *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Synthese Library* (Vol. 359). Cham: Springer, 79-94.

Warglien, Massimo, Peter Gärdenfors, and Matthijs Westera. 2012. Event structure, conceptual spaces and the semantics of verbs. *Theoretical Linguistics* 38: 159-193.

Watanabe, Satosi. 1969. *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: John Wiley.

Webb, Andrew R. 2002. *Statistical Pattern Recognition* (2nd edition). Chichester: John Wiley.

Weiskopf, Daniel A. 2009a. The plurality of concepts. *Synthese* 169: 145-173.

—. 2009b. Atomism, pluralism, and conceptual content. *Philosophy and Phenomenological Research* 79: 131-163.

—. 2010. The theoretical indispensability of concepts. *Behavioral and Brain Sciences* 33: 228-229.

Werning, Markus. 2005. Right and wrong reasons for compositionality. In M. Werning, E. Machery and G. Schurz, eds., *The Compositionality of Meaning and Content* (Vol. I). *Foundational Issues*. Frankfurt: Ontos Verlag, 285-309.

Wiener-Ehrlich, Willa K., William M. Bart, and Richard Millward. 1980. An analysis of generative representation systems. *Journal of Mathematical Psychology* 21: 219-246.

Wilkenfeld, Daniel A. Forthcoming. Understanding as compression. *Philosophical Studies*, https://doi.org/10. 1007/s11098-018-1152-1.

Wilmink, Frederik W., and Hilde T. Uytterschaut. 1984. Cluster analysis, history, theory and applications. In G.N. Van Vark and W.W. Howells, eds., *Multivariate Statistical Methods in Physical Anthropology: A Review of Recent Advances and Current Developments*. Dordrecht: Reidel Publishing Company, 135-175.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*, C.K. Ogden (trans.), London: Routledge & Kegan Paul. Originally published as "Logisch-Philosophische Abhandlung", in *Annalen der Naturphilosophische* (Vol. XIV, 3/4), 1921.

—. 1953. *Philosophical Investigations*, G.E.M. Anscombe and R. Rhees, eds., G.E.M. Anscombe, trans. Oxford: Blackwell.

Yates, Allen. 1987. *Multivariate Exploratory Data Analysis: A Perspective on Exploratory Factor Analysis*. Albany, NY: State University of New York Press.

Yu, Chen, Linda B. Smith, Krystal A. Klein, and Richard M. Shiffrin. 2007. Hypothesis testing and associative learning in cross-situational word learning: Are they one and the same? In E.S. McNamara and J.G. Trafton (eds.), *Proceedings of the 29th Cognitive Science Society Conference*. Mahwah, NJ: Lawrence Erlbaum Associates, 737-742.

Zadeh, Lotfi A. 1965. Fuzzy sets. *Information and Control* 8: 338-353.

Zalta, Edward N. 2001. Fregean senses, modes of presentation, and concepts. *Philosophical Perspectives* 15: 335-359.

Zenker, Frank, and Peter Gärdenfors. 2015a. Communication, rationality, and conceptual changes in scientific theories. In F. Zenker and P. Gärdenfors, eds., *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Synthese Library* (Vol. 359). Cham: Springer, 259-277.

—. 2015b. *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Synthese Library* (Vol. 359). Cham: Springer.

Zhang, Siyu, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo, and Yang Dan. 2014. Long-range and local circuits for top-down modulation of visual cortex processing. *Science* 345: 660-665.

Zielman, Berrie, and Willem J. Heiser. 1996. Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology* 49: 127-146.

Zubin, Joseph. 1938. A technique for measuring like-mindedness. *Journal of Abnormal and Social Psychology* 33: 508-516.