University of Trento

CIMeC – Centre for Mind/Brain Sciences

Doctoral School in Cognitive and Brain Sciences (XXXI Cycle)

Track Cognitive Neuroscience

FROM PERCEPTUAL TO SEMANTIC REPRESENTATIONS IN THE HUMAN BRAIN

Simone Viganò

Supervisor: prof. Manuela Piazza

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive and Brain Sciences

March 2019

ABSTRACT

Humans are capable of recognizing a myriad of objects in everyday life. To do that, they have evolved the ability to detect their commonalities and differences, moving from perceptual details to construct more abstract representations that we call *concepts*, which span entire categories (such as the one of people) or refer to very specific and individual entities (such as our parents). Organizing our knowledge of the world around concepts, rather than around individual experiences, allows us for more rapid access to behavioural relevant information (for instance, how to behave when we encounter a dangerous animal), and to quickly generalize this information to what we never encountered before. In few words, this is what permeates the world with meaning.

The present work is about the neural bases of learning novel object concepts, a process that in our species is vastly supported by symbols and language: for this reason, I talk about *semantic* representations. The word "semantics" generally refers to the study of meaning (and to what a "meaning" ultimately is) as it is conveyed by a symbol; in the specific case of cognitive neuroscience, it deals with the neural mechanisms that allow symbols to *re-present* the meanings or concepts they refer to in the brain. For instance, we can easily describe what is the meaning of the word "DOG", pretty much as we can explain what "DEMOCRACY" *means*. However, although cognitive neuroscience has focused on the neuro-cognitive bases of semantic representations for decades, the neural mechanisms underlying their acquisition remain elusive. How does the human brain change when learning novel concepts using symbols? How does a symbol acquire its meaning in the brain? Does this learning generate novel neural representations and/or does it modify pre-existing ones? What internal representational format (neural code) supports the representation of newly learnt concepts in the human brain?

The contribution of this work is three-fold. First, I show how new semantic representations learned by categorizing novel objects (defined through a combination of multisensory perceptual features)

using words, emerge from the orchestrated plasticity of both perceptual and memory systems. Second, I show results converging on the idea that brain regions that evolved in lower-level mammals to represent spatial relationships between objects and locations, such as the hippocampal formation and medial prefrontal cortex, in humans are recruited to encode relationships between words and concepts by means of the same neural codes used to represent and navigate the physical environment. Finally, I present preliminary data on the cognitive effects of using symbols during learning novel object concepts, showing how language supports the construction of generalizable semantic representations.

ACKNOWLEDGMENTS

Thanks to Manuela, for being an advisor, a guide, a mentor.

Thanks to my friends, parents, and relatives.

Thanks to Andrea, *perché ho bisogno della tua presenza*.

The universe (which others call the Library) is composed of an indefinite and perhaps infinite number of hexagonal galleries, with vast airshafts between, surrounded by very low railings. […] Like all men of the Library, I have travelled in my youth; I have wandered in search of a book, perhaps the catalogue of catalogues; now that my eyes can hardly decipher what I write, I am preparing to die just a few leagues from the hexagon in which I was born.

Jorge Luis Borges (1941), *The Library of Babel*, in *The Garden of Forking Paths*

TABLE OF CONTENTS

GENERAL INTRODUCTION

A key step for making sense of the rich multisensory world surrounding us is to be able to parse it into meaningful discrete and recognizable object categories, or "concepts". To solve this task, the human brain needs to extract from experience and combine all the defining details of a concept, such as its sensory or contextual properties. This set of information we have about things in the world is called "*conceptual knowledge"*, and it defines the bases for nearly every human activity: it allows us, for instance, to remember what distinguishes dogs from wolves, how to use a pen to write, or how to behave in a particular situation. A long-standing tradition in cognitive neuroscience has referred to the neural representations of concepts in the human brain as *semantic* representations, because of the central and undeniable role that language and symbols (such as words or numbers) play in acquiring, organizing, and recalling conceptual knowledge (semàntico derives from the late latin semantĭcus, and from the greek term σημαντικός - "meaningful", from the root σημαίνω «to symbolize, to mean»). The present work is about how semantic representations emerge in the human brain, how they are organized, and what are their effects on human behaviour.

This thesis is divided into 5 chapters. Chapter 1 is an introduction to the neuroscience of semantic representations, and it will revolve around two main themes: the representation and acquisition of concepts in the human brain through symbols (an issue known as the "symbol-grounding problem" (Harnad 1980)), and the neural codes supporting the organisation of these representations. Bringing together knowledge from previous studies, I will formulate two predictions: 1) that novel semantic representations, defining the meaning of symbols (words) emerge in the brain as a consequence of the orchestrated plasticity of both memory and perceptual systems, and 2) that in humans the same neural codes subtending spatial navigation, might also support the representation of language-based semantic knowledge, especially in those brain regions mostly known for encoding spatial memory and higher level reasoning, such as the hippocampal

formation and the medial prefrontal cortex. By the end of Chapter 1 I will present a behavioural training paradigm thanks to which human adults learn novel object concepts that I validated and used for subsequent experiments. In Chapter 2 and 3 I will present a longitudinal fMRI study I designed to test these predictions. In this study, participants were trained for 9 days to construct a novel semantic space, and crucially they were tested before and after this learning process. I will present the two sets of analyses separately, in the format of two independent journal articles, as they attack two questions that I believe(d) being distinguishable. Chapter 2 will summarize the neuroimaging results addressing the first prediction. Chapter 3 will address the idea that spatial neural codes (distance-based and direction-based) employed by specific brain regions to represent the structure of the physical space and to support spatial navigation, may also be recruited to represent the structure of the novel semantic space during an orthogonal and non-spatial symbolic categorization task. In Chapter 4 I will present the results of an independent behavioural investigation aimed at verifying the advantages of using symbols to create novel categorical representations: although at the moment of writing this experiment is still ongoing, I believe the first effects that it shows will be of particular interest for the present dissertation. Finally, in Chapter 5 I will conclude this work with a general discussion, stressing open questions and future directions.

Semantic representations in the human brain

The nature of concepts has been a central topic in philosophy and cognitive science for centuries. A long standing tradition in cognitive neuroscience attacked the question of the neural correlates of human conceptual knowledge by taking advantage of the fact that in humans conceptual memory is dependent on symbols, such as words or numbers, and can be inferred by mostly using linguistic material. Such symbol-dependent form of conceptual knowledge is defined "semantic memory" and its study aims at unveiling the neuro-cognitive mechanisms that give rise to semantic representations.

Early neuropsychological studies (e.g. Warrington & McCarthy 1983; Warrington and Shallice 1984) indicated that brain damaged patients have selective deficits for some categories of concepts compared to others. More than one hundred cases has been reported so far (for reviews see Capitani et al. 2003; Mahon & Caramazza 2010), involving semantic specific impairments for living things such as animals (e.g. Caramzza & Shelton 1998; Blundo et al. 2006), fruit/vegetables (e.g. Hart et al. 1985; Samson & Pillon 2003), non-living things such as tools (e.g. Laiacona & Capitani 2001; Sacchett & Humphreys 1992); and conspecifics (e.g. Ellis et al. 1989; Miceli et al. 2000).

Other studies tried to further extend these results by mapping what neocortical regions were activated when healthy participants were engaged in various conceptual tasks prompted by words presentation. The rationale behind this approach is that words, as abstract symbols provided with meanings by experience, guarantee a rapid and efficient access to conceptual representations while at the same time controlling the contribution of low-level properties of the physical input. When participants are asked to perform tasks that enhance the meaningful nature of words, as opposed for instance to pseudowords (e.g. Demonet et al 1992; Binder et al. 2003, 2005; etc etc.), only those neocortical regions that store and elaborate purely conceptual information should activate: the semantic network.

Despite the high variability in the kind of words (e.g. concrete vs. abstract words) or tasks (e.g. evaluate if a string of letters was a words or a pseudo-words, or whether two words referred to similar concepts or not) used, these studies generated very consistent results. In 2009 Binder et al. published a critical review and meta-analysis of 120 functional neuroimaging (fMRI and Positron Emission Tomography, or PET) studies focusing on conceptual processing on healthy adults. These studies were conducted in laboratories all over the world in a period of time that span more than 15 years (from 1992 to 2007), and involved a great variety of conceptual tasks. The results revealed a distributed, but mainly left-lateralized network of 7 cortical regions consistently activated: the left Angular Gyrus in the inferior parietal cortex; 2) middle and inferior temporal gyri extending to the anterior temporal lobe; 3) fusiform and parahippocampal gyri; 4) dorsomedial prefrontal cortex; 5) ventromedial prefrontal cortex; 6) inferior frontal gyrus; 7) precuneus and posterior cingulate cortex (Figure 1.3). All the nodes of this network are associative regions far from primary sensory and motor cortices, and they are consistently reported as high-level multimodal areas (Mesulam 1985; Sepulcre et al. 2012) with wide and distributed connectivity with secondary sensory areas (Achard 2006; Buckner et al. 2009; Andersen et al. 1990).

In 2013, Fairhall & Caramazza directly investigated what brain regions showed the definitional property of neural semantic representations of object concepts (that is of concepts that represent object classes) – that of showing corresponding, or similar, activation patterns for a concept, irrespective to its presentation modality (either the symbol (e.g., the word CAT) or its referent (e.g., the picture of a cat) – by using multivariate analysis to identify neocortical regions that represented well known object concepts in a modality invariant fashion. They presented participants with stimuli belonging to 5 semantic categories – fruits, clothes, tools, mammals, and birds – during a typicality-judgment task (e.g. rating the typicality of “hammer” as a tool, or “apple” as a fruit). Crucially, participants were presented with these stimuli in either a pictorial (e.g. the picture of an apple) or a symbolic (e.g. the string of letters A-P-P-L-E) format. The authors then applied a cross-modal decoding procedure implemented in a whole-brain searchlight: for each sphere of the searchlight, a classifier (SVM) was trained to distinguish semantic content from the multivariate

brain activity evoked during the presentation of stimuli in one sensory modality (e.g. pictures), and it was tested on the independent brain activity dataset collected during stimuli presentation in the other modality (e.g. words). This revealed a network of areas mostly overlapping to the semantic network, thus indicating that these regions are indeed representing semantic content irrespective of the format (symbolic or pictorial) of presentation.

Among these neuroimaging studies using words to access semantic representations, the most consistently reported neocortical region is the left Angular Gyrus (AG), in the inferior portion of the parietal lobule (IPL). This region is practically absent in lower primates, and expanded significantly in humans compared to their homologues in macaques (von Bonin and Bailey 1947; Hyvarinen 1982). Given its anatomical location at the conjunction of secondary visual, auditory, spatial, and somatosensory associative regions, it has been indicated as the ideal candidate as neo-cortical "convergence zone" (for a definition see Damasio 1989 and Meyer & Damasio 2009), where high-level conjunction of perceptual information is integrated into more abstract, or conceptual, representations (Geschwind 1965; Binder & Desai 2011). This view has been confirmed by brain stimulation studies, causally linking modulation of AG activity to modulation of behaviour and performance in memory tasks (e.g. Sestieri et al. 2013). Yazar et al. (2017) applied continuous theta burst stimulation to this area while participants had to retrieve information on audio-visual features of recently acquired memories. They showed a significant impairment in participants when they had to retrieve conjunctive multisensory information (audio and video together) compared to a condition where the stimulation was applied at a vertex control site, and no effect when they had to retrieve single modality features (audio or video separately). This indicated a specific role of the Angular Gyrus in combining (or binding) multisensory information, an operation that in real life is essential for defining new memories and concepts, and also for grounding the meaning of new symbols.

However, other works mostly based on clinical observations, indicated the anterior portion of the left temporal lobe (ATL) as the key convergence hub for semantic processing (for reviews see Patterson et al. 2007; Lambon-Ralph 2014). Crucial evidence in this sense comes from a dramatic

neurodegenerative disorder, semantic Dementia (SD). Mostly affecting temporal regions, this disorder is characterized by severe anomia and inability to recover conceptual knowledge even in tasks that do not require its explicit verbal communication, such as simple object use (Hodges et al 2000) or item identification based on sound (Bozeat et al. 2000), taste (Piwnica-Worms et al 2010) or smell (Luzzi et al. 2007). SD patients are usually impaired in judging the typicality of items within a conceptual category (e.g. guitar as musical instrument), and their performance decreases as a function of specificity (e.g. recognizing a very specific dog breed)(Lambon Ralph et al. 2016). Neuroimaging studies confirmed that anterior regions of the temporal lobe differently represent concepts on the bases of their semantic details, such as categorical membership (Malone et al. 2016, Borghesani et al. 2016).

All these studies suggest that processing semantic knowledge in humans elicits activity in a widespread network of associative regions that presumably, in light of their specific anatomical positions, act as convergence zones (Meyer & Damasio 2009) for inputs coming from lower associative and sensory regions.

But how do these semantic representations emerge in the brain? A fundamental problem in cognitive science, indeed, is the “symbol-grounding problem” (Harnad 1980), that relates to the issue of how a symbol acquires its meaning. In the field of cognitive neuroscience, this translates to the question of whether and how the neural representations of symbols and the objects they refer to change to reflect the novel, meaningful, association, or whether this novel representation emerge separately and independently in brain regions that did not previously represent either the symbols or the objects themselves.

The observation that in some cases there are shared neural representations between a symbol and its non-symbolic meaning suggests that symbols acquire their meaning by means of a mapping process onto the same neural representation of their referent (Pulvermuller 2013). This seems particularly true in the case of numbers, where brain regions responding to quantity, such as the Intraparietal Sulcus (hIPS), show a representational code common to both number symbols and non-symbolic numerosities, as revealed by fMRI adaptation (Piazza et al. 2007), to the point

that even the semantics of complex mathematical sentences activates the same neuronal circuits usually involved in processing simpler numerical operations or over digits but also sets of items (Amalric & Dehaene 2016).

However, as these studies focused on well known semantic categories, it is not possible to have a conclusive answer, neither to unveil what are the brain mechanisms that allow this putative mapping or grounding process to happen: indeed, these results are silent on whether the computations necessary to attach a symbol to its meaning (and viceversa) happen within the same areas the later show the mutual correspondence, or if other areas participates in building the novel semantic representation.

One potentially powerful way to address the problem is to use training studies, where participants learn new concepts by associating them with specific names. The use of functional neuroimaging techniques then permits to record the activity patterns for the stimuli at different time points, for instance before and after learning the semantic association, and thus reveal what are the changes occurring in the brain as new meanings are created.

The behavioural consequences of learning novel with the use of symbols compared to learning it without symbolic aids have been indeed explored in behavioural training experiments showing that, for instance, the availability of symbols greatly facilitates the acquisition of novel categories both in adult (Lupyan et al. 2007) and children (e.g. Althaus & Plunkett 2016, Althaus & Westemann 2015). There are reasons, coming from behavioural studies, to believe that the changes supporting the emergence of semantic information spread also to perceptual representations. Past works, indeed, highlighted the effects of categorization on perceptual judgements. Long-lasting expertise can improve perception of diagnostic structures and features in animal (Biederman & Shiffrar 1987) or beer (Peron & Allen 1988) experts, as well as in radiologists (Norman et al. 1992), suggesting that learning to recognize specific object categories by attaching them a label can alter perceptual processing. Such a categorical effect on perceptual judgements seems to rely on dimensional modulation (Folstein et al. 2015) of behaviourally relevant perceptual features. This

alteration revolves around acquired distinctiveness between members of different categories (Lawrence 1949) and acquired equivalence between members of the same category (for a review see Braunitzer et al. 2017). Goldstone (1994), for instance, found that participants who have been trained to categorize, using labels, a set of 16 squares basing on their size and brightness were more likely to discriminate between across- boundary stimuli compared to a control group, providing behavioural evidence for acquired distinctiveness. This "warping" of the perceptual representations occurring during categorization might indicate that learning new semantic knowledge involves changes that may be traced down to the perceptual systems as well, effects that are usually overlooked by studies that focused on well-known classes of object/words.

This might indicate that the symbol-grounding problem is solved by the human brain by means of complex and distributed changes that spread even to perceptual representations.

Conclusions

Humans construct their conceptual knowledge of the world by organizing multisensory experiences into labelled categories. No study to date systematically looked for the neural changes supporting this uniquely human faculty monitoring the early stages of learning to map symbols to their meaning. Several questions remain open: does learning generate *ex-novo* neuronal representations that were not present before? Does it also, or only, modify previously existing ones? Do these changes involve brain regions beyond the semantic network? What neural mechanisms support these changes?

To answer these questions, I designed a learning experiment where I monitored, using fMRI, the neural changes of learning novel multisensory object concepts using symbols. The details of the experiment will be presented in Chapter 2, while the remaining part of this introductory chapter will revolve around the second central topic of the present work: what is the representational format underlying semantic representations?

Spatial codes for semantic spaces

The second part of this introduction is about the format underlying conceptual representations in the human brain. This refers to the neural code(s) that different brain regions employ to represent the relations between concepts. Specifically, I will discuss a fascinating idea that emerged in the late 40s by Tolman (1949) and that has been recently formalized in a theoretical work (Bellmund et al. 2018), suggesting that the representation of the knowledge we have about things in the world and that we use in our everyday behaviour is supported by the same neural mechanisms that we recruit to represent the physical space. This theory states that the relationships between concepts and items in memory are conceivable as distances between the regions of a conceptual representational space, and thus we can use the same neural codes that allow us to navigate in the physical space (spatial codes) to "move" among concepts in memory.

Between the 30s and the 40s Tolman conducted a series of behavioural experiments on rats, where he observed that animals, to find rewards in complex mazes, were able to take shortcuts or find new routes when the old ones were blocked (e.g. Tolan & Honzik 1930, Tolman et al. 1946). He coined the term "cognitive map" to indicate that the animals, in order to show such complex and adaptive behaviour, must have had developed an internal representation of the world and the relationships between its elements, such as landmarks or locations (Tolman 1948).

A literal interpretation of the word "map" directed the following years of research to find the internal neural correlates of such representation of the external physical environment. In 1971 O'Keefe and Dostrovsky discovered hippocampal "place cells", neurons that are active when the animal enters very specific positions in the environment, no matter the orientation of the movement trajectory or its velocity. The following four decades have seen a proliferation of milestone results in the study of spatial coding in this area, mostly represented by the discovery of other spatially tuned neurons,

such as head direction cells (Ranck 1984; Taube et al 1990), boundary cells (O'Keefe and Burgess 1996), boundary vector cells (Lever et al. 2009), speed cells (Kroppf et al. 2015), object vector cells (Hoyadal et al. 2017) and most recently even social place cells (Omer et al. 2018; Danjo et al. 2018).

The most celebrated kind of spatially tuned neurons are grid cells, first described by the group of Edvard and Mary-Britt Moser (hafting et al. 2005), who in 2014 were awarded, together with John O'Keefe, the Nobel Prize in Medicine and Physiology. Grid cells were first observed in the medial entorhinal cortex of rats (a sub-portion of the hippocampal formation, mostly projecting to the hippocampus), and are neurons that fire for multiple spatial locations in the environment. These locations correspond to the vertices of a regular triangular grid covering the entire environment, and show a precise 6-fold rotational symmetry, resulting in a very specific hexagonal pattern. Besides their peculiar firing rate, grid cells show some other very interesting properties. First, visual cues strongly influence the alignment of the grid: when external cues are rotated, the grid pattern rotates in the same way. Second, grid activity remains unchanged when visual input is removed (e.g. by turning off the lights in the environment). Third, grid patterns appear as soon as the animal enters a novel environment. Finally, and possibly most importantly, grid cells maintain the specific size of the grid pattern and its offset compared to one another even if the animal is moved to different environment. This property is not shown for instance by hippocampal place cells, that exhibits a profound remapping in different environments (Bostock et al. 1991; Leutgeb et al. 2005; Fyhn et al. 2007). In general, grid cells are thought to support path-integration, enabling an internal representation of distances between locations, thus guiding mammals' behaviour when navigating the environment (Bush et al 2015).

A seminal study by Doeller et al. (2010) demonstrated that grid activity is present in humans, and that it is possible to record it using non-invasive functional MRI. This study moved from a very precise observation about electrophysiological data on rats: grid orientation of different grid cells relative to the external environment remains constant across cells (while for instance their relative phase or size of the grid pattern change). To understand why this observation is so important,

consider a single grid cell, which activates more often when the animal moves in the environment in a direction that is aligned to one of the 6 main axes of the grid, compared to a situation where it moves for the same distance but in a direction that is not aligned. If we consider now entorhinal activity at the population level, this would result in a stronger signal for movement directions aligned to the grid (one of the 6 axes) compared to movement directions that are not aligned to the grid.

The brilliant intuition of the authors was that such different population activity should require a different consumption of blood, thus it could be observed at cortical level as a modulation of the BOLD signal, using functional MRI. Doeller et al. in their experiment asked participants to navigate a virtual reality environment with a joystick, while lying in the MR scanner. During navigation, participants had to find the locations of some objects, while their brain activity was analysed, looking for 6-fold modulations of the BOLD signal as a function of running direction (at this step, randomly aligned to a reference direction). The analysis technique they used was particularly complex, and consisted in two steps. In a first step, half of the functional data were used to estimate the putative grid orientation, by means of a quadrature filter procedure. Next, they aligned the running directions of the second, and independent, half of the dataset to the putative grid orientation, and looked for intensity of the BOLD signal for aligned vs. misaligned clusters (30°) of movement directions. They reported an impressively precise modulation of BOLD signal in the right entorhinal cortex, that could not be explained by other periodicities (e.g. 45° or 90°). Crucially, when they applied fMRI adaptation to reveal those brain regions that showed a reduction in fMRI signal according to how recently participants were running at 60° to the current direction, this revealed a network of areas including not only the entorhinal cortex, but also other areas, such as the medial prefrontal cortex, best known for its connectivity to the hippocampal formation and for its role in both spatial e non spatial memory (Preston & Eichenbaum 2013).

In 2013 Jacobs et al. reported the first evidence of grid cells in humans using intracranial recordings, while epileptic patients performed a virtual reality task. Neurons in their entorhinal

cortex and in medial prefrontal cortex exhibited grid-like firing patterns as a function of spatial position in the virtual environment, thus proving that humans and lower level animals rely on corresponding spatial-coding schemes at neuronal level. Interestingly, two independent studies in 2016 observed grid-like modulation of fMRI BOLD signal when healthy participants were involved in imagined navigation tasks. In the first of these studies, Horner et al. (2016) trained participants to memorize the positions of 6 objects in a virtual reality environment. Next, they asked them to either move or imagine moving to the locations of each object, from various positions, thus eliciting different movement trajectories. ROI-based analysis revealed a significant cluster of voxels in EC that showed 6-fold modulation of bold signal as a function of running direction. In the second study, Bellmund et al. (2016) independently confirmed these results by applying a more parsimonious and potentially powerful method based on Representational Similarity Analysis (RSA, Kriegeskorte et al. 2008), where they showed that the neural similarity of pairs of imagined movement trajectories – carefully sampled to be at 30° or 60° apart one from each other – was higher, in EC, when the two trajectories were 60° apart compared to when they were 30° apart, as an underlying grid-code would impose.

As both the hippocampal formation and the medial prefrontal cortex are classically associated to more general memory functions (see Preston & Eichenbaum 2013, Stalnaker et al. 2015; Behrens et al. 2018 for reviews), is it possible that the same spatial codes are involved in non-spatial navigation tasks? Constantinescu et al. have made a crucial contribution in this sense in 2016. They adapted the same logic and experimental design of Doeller et al. (2010) to ask whether the same grid-like activity could be observed, using fMRI, when participants processed a novel continuous space of visual shapes. They created 6 bird shapes and they associated each one of them to a Christmas symbol. Crucially, bird shapes varied in the length of their legs and neck, thus each bird could be intended as a point in a bi-dimensional “bird” space where coordinates were the length of the two diagnostic features. They made participants familiarize with this bird space by means of a task where they could adjust the ratio between neck-length and legs-length, thus mimicking a movement in this artificial space. By adjusting these two visual features, participants

had to find the 6 birds shapes associated to the Christmas symbols. Next, during the fMRI sessions, participants were presented with brief videos of morphing birds, showing a slow change in their silhouette in terms of neck- and legs-length. Participants were instructed to imagine the morphing animation to continue “in the same way” (that is, crucially, in the same direction in the corresponding 2D bird space) and to guess what kind of resulting bird shape they will find, as indicated by one of the Christmas symbols. Although participants were not consciously aware of the 2D spatial representation underlying this task, when authors analysed their brain activity as a function of “morphing” direction looking for the 6-fold periodic modulation typical of grid-cells activity, they found it in a network of areas strikingly similar to the one reported by Doeller et al. (2010) for spatial navigation. In particular, this signal was stronger in the entorhinal cortex and in the ventromedial prefrontal cortex.

This result was the first, and to date the only one, evidence of hexadirectional modulation for a non-spatial task in humans, which required memorizing a continuous and bi-dimensional visual space. This proves that the grid-code might serve, in the human brain, a more general function than representing the physical space, and it opens the possibility of representing conceptual knowledge using spatial codes.

But what does it mean to represent knowledge using spatial codes? In the theoretical framework formalized by Gardenfors (2000), knowledge can be conceived as organized into “cognitive spaces”, internal representations of objects or events spanning by a set of quality dimensions (sensory or abstract features). For instance, a zebra and a wasp can be thought as occupying different regions in a bi-dimensional “animal space” spanning animals’ size and ferocity, or any other two dimensions might be relevant for the task to solve or for the memory to encode. Any given stimulus can be thus located in a cognitive space according to a set of diagnostic feature values. Relations between concepts (regions of the cognitive space) can be expressed using geometrical notions: dissimilarity between concepts can be expressed as Euclidean distance between regions in the n-dimensional feature space, and sequences of concepts are thus

conceivable as movements in the corresponding underlying space. Interestingly, a very similar intuition is also emerging in completely different fields, that of neurolinguistics and computational linguistics, where scholars tend to conceptualize word meanings (that is, semantic representations of concepts), as regions or points in an internal space, the semantic space, with proximities reflecting similarity in meaning, thus highlighting that high-level symbolic thinking might share some important features with spatial processing (Borghesani & Piazza 2017). Under this framework, it is essential to provide an interface to index the location of a concept along one or more dimensions. Place- and grid- cells do that for physical spaces, easily conceivable as bi-dimensional navigable surfaces, but they might serve the same purpose for any conceptual representation that can be reduced to an n-dimensional space of task relevant features.

Consider the study by Aronov et al. (2017). In the task they designed, rats were required to use a joystick to manipulate a sound along a 1-dimensional continuous frequency axis, to find the correct frequency that would lead to a reward. They recorded neural activity in the hippocampus and in the entorhinal cortex, and they found that both regions contained neurons that responded to very specific sound frequencies. In particular, neurons in the hippocampus fired selectively for only one frequency each, while neurons in the entorhinal cortex exhibited multiple firing fields at different (usually 2-3) sound frequencies. Crucially, to test whether these neurons were also involved in spatial representations, they recorded their activity while rats navigated a spatial environment looking for pellets of food. They found that between 25% and 35% of spatially tuned cells were also involved in the sound modulation task. These results indicate that during a non-spatial task, the hippocampal-entorhinal system of lower-level mammals holds a representation of the task relevant features (in this case just one, sound) in a 1-dimensional feature space, where different regions or states (the frequencies) are represented by the same neurons that represents locations in the physical environment, showing similar firing properties (e.g. single selective vs. multiple firing fields for place and grid cells, respectively). As spatial and non-spatial task representations are produced by the same neuronal population, the underlying neural code(s) – usually referred to as spatial code(s) in light of their first observation in spatial tasks - might serve a more general

function, such as representing the underlying structure of an internal representation of the task: exactly what Tolman called “cognitive map” and what Gardenfors called “cognitive space”.

Conclusions

Humans and lower level mammals rely on the same neural mechanisms to navigate the physical space, recruiting a variety of spatial codes mostly encoded in the hippocampal formation. However, the same spatial-codes that allow to navigate the physical space have been observed in humans in non-spatial tasks, such as evaluating visual shapes corresponding to regions of a perceptual bi-dimensional visual space. This suggests that in humans, the same structures and neural codes that subtend spatial representations might also be recruited for more abstract and higher-level forms of cognition. To date, no study has investigated more thoroughly this intuition. Do “spatial” codes activate to represent human semantic knowledge, which is multisensory, categorical, and highly dependent, by definition, on symbols and language?

I will address this point specifically in Chapter 3, where I will use multivariate analysis to explore the existence of both a distance and a direction-based code of a novel semantic space during a symbolic categorization task.

Introduction to the experimental work

In the next chapters I will describe three works trying to attack 3 fundamental questions in the study of semantic representations:

1. how do semantic representations emerge in the human brain?
2. does the human brain recruit spatial codes for representing semantic information even when it has no spatial content?
3. does learning categories of objects using symbols facilitate generalization to novel exemplars?

The first two works will describe a set of longitudinal fMRI analyses combined with a 9-days long symbolic categorical training, that represented the core of my work during this doctoral program. The third work, which is still ongoing, will present the very preliminary, yet of potential interest, results of a behavioural investigation.

The first part of my doctorate has been dedicated, besides the study of the relevant literature and of the neuroimaging methods that I will be describing later on, to validate a behavioural training paradigm suitable for later experiments. This revolved around i) the creation of a novel semantic space composed by multisensory objects, which are divided into 4 orthogonal categories by means of abstract labels (novel words), and ii) the validation of the behavioural training. I will briefly sum up the methods and the results of this validation as final part of this introductory chapter, before moving to the presentation and the discussion of the experimental work.

Participants. The study included 15 right-handed adult volunteers (10 females and 5 males; mean age = 21.6, std = 2.02). All participants gave written informed consent and were reimbursed for their time.

Stimulus space. I developed a set of 16 novel animated multisensory objects, orthogonally manipulating the size of an abstract shape (Figure 1.1A) and the pitch of an associated sound. A total of four size- and pitch- levels were used for each participant, leading to a stimulus space where each object represented the unique combination of one size and one pitch level (Figure 1.1B). The values of these two features were selected for each participant on the first day of the experiment, following a brief psychophysical validation consisting of a QUEST adaptive staircase method (Watson & Pelli 1987). Using a two-stimuli comparison task for each sensory modality, I calculated subject-specific sensitivity as the minimum appreciable increment (Just Noticeable Difference, JND) from a reference value (size: visual angle of 5.73°, pitch: frequency of 800 Hz) leading to 80% of correct responses. For each sensory modality, four subject-specific feature levels were calculated, applying the logarithmic Weber-Fechner's law and selecting values at every three JNDs, in order to ensure that feature levels were equally distant and clearly identifiable. Moreover, in order to strengthen the multisensory binding between the two unisensory features, I applied a 'squeezing' animation during each object presentation by displaying 13 frames of the same object with increasing (frames 1 to 7) and decreasing (frames 8 to 13) size along the horizontal axis (for an exemplar video of the animated stimuli, visit https://www.youtube.com/watch?v=Nyq2BgY-8jc&feature=youtu.be). Objects presentation lasted a total of 750 ms and sounds were presented at the apex of the squeezing period. The object space was divided into four categories based on the combination of two sensory boundaries (Figure 1.1B). The categorical membership of each object, as well as their unique multisensory identities, could thus be recovered only when considering both sensory features. I assigned to each category an abstract name (Figure 1.1C): KER (small size and low pitch); MOS (big size and low pitch); DUN (small size and high pitch); GAL (big size and high pitch).

Stimuli presentation. Stimuli were presented foveally using MATLAB Psychtoolbox in all experimental phases, at a distance of ~ 130 cm. Multisensory objects subtended a different visual angle for each size level, and a different frequency for each pitch level, ranging from an average of 5.73° and 800 Hz for level 1 (size and pitch, respectively) to an average of 8.97° and 973.43 Hz for

level 4. Each word subtended a visual angle of 3.58° horizontally and 2.15° vertically, and was presented with black Helvetica font on a grey background.

Stimuli presentation. Stimuli were presented foveally using MATLAB Psychtoolbox in all experimental phases, at a distance of ~ 130 cm. Multisensory objects subtended a different visual angle for each size level, and a different frequency for each pitch level, ranging from an average of 5.73° and 800 Hz for level 1 (size and pitch, respectively) to an average of 8.97° and 973.43 Hz for level 4. Each word subtended a visual angle of 3.58° horizontally and 2.15° vertically, and was presented with black Helvetica font on a grey background.

Behavioral training. Participants underwent 9 daily sessions of behavioral training, aimed at making them learn the correct name of each object. Each behavioral session was approximately 10 minutes long, and it was divided into 4 mini-blocks of 20 trials each, for a total of 80 trials. It started with a brief presentation of the objects as exemplars of the four categories (KER, MOS, DUN, GAL). After this familiarization phase, each trial consisted of an object presentation (750 ms), followed by a fixation cross (500 ms), and by the presentation of the 4 possible names in random order, from left to right (Figure 1.1D). Each object was presented 10 times per training session. Participants were instructed to press one key on the keyboard to select the correct name. They were asked to respond as fast as possible, but no time limits were imposed. After their response, an immediate feedback appeared on the screen for 1000 ms, indicating with the words “Correct!” or “Wrong!” the accuracy of the choice. In the case of a wrong answer, the feedback also showed the correct object name, in order to speed up the learning process. After each miniblock, participants would be provided with the cumulative percentage accuracy. Starting from the seventh training session, the trial-by-trial feedback was removed and participants could rely only on the block-by-block cumulative feedback. For the first 8 days of training, participants were presented with the same 8 objects used in the two fMRI sessions. On the last training day, without being notified of the change, they were presented with all 16 objects. This allowed me to test for generalization of the categorical rule to new exemplars (here represented by objects 2 - 4 - 5 - 7 -

10 - 12 - 13 - 15), which would be a key ingredient of an efficient semantic representation. For this last session, the mini-blocks number was kept at 4, but the number of trials was doubled, resulting in a total testing time of ~ 20 min.

Behavioral training results.The learning trajectory indicated an increment in performance from session 1 to 8 (session 1: 46.5 ± 16%; session 2: 66.33 ± 20%; session 3: 75 ± 18%; session 4: 77.5 ± 15%; session 5: 78.6 ± 19%; session 6: 78.7 ± 18%; session 7: 79 ± 14%; session 8: 78 ± 19%; difference from session 8 to session 1: t_{14} = 7.25, p = $4.17x10^{-6}$). Performance collected on session 9 confirmed both the successful learning of the name-objects association and its generalization (training set: mean accuracy = 79.5%, std = 17%, different from chance t = 16.71, p < .001; generalization set: mean accuracy = 75.58%, std = 15.29%, difference from chance, t = 12.81, p < .001)(Figure 1.1E).

Conclusions

The results of this first behavioural validation indicates that participants correctly learned to categories the novel multisensory space using words. Moreover, by analysing the performance on the last training day (test day), I can conclude that the categorical meaning of the novel word was acquired in an abstract and generalisable way, because participants could correctly categorise novel exemplars, a key ingredient when creating behaviourally relevant semantic representations. This behavioural paradigm will be largely used in the following experimental works.

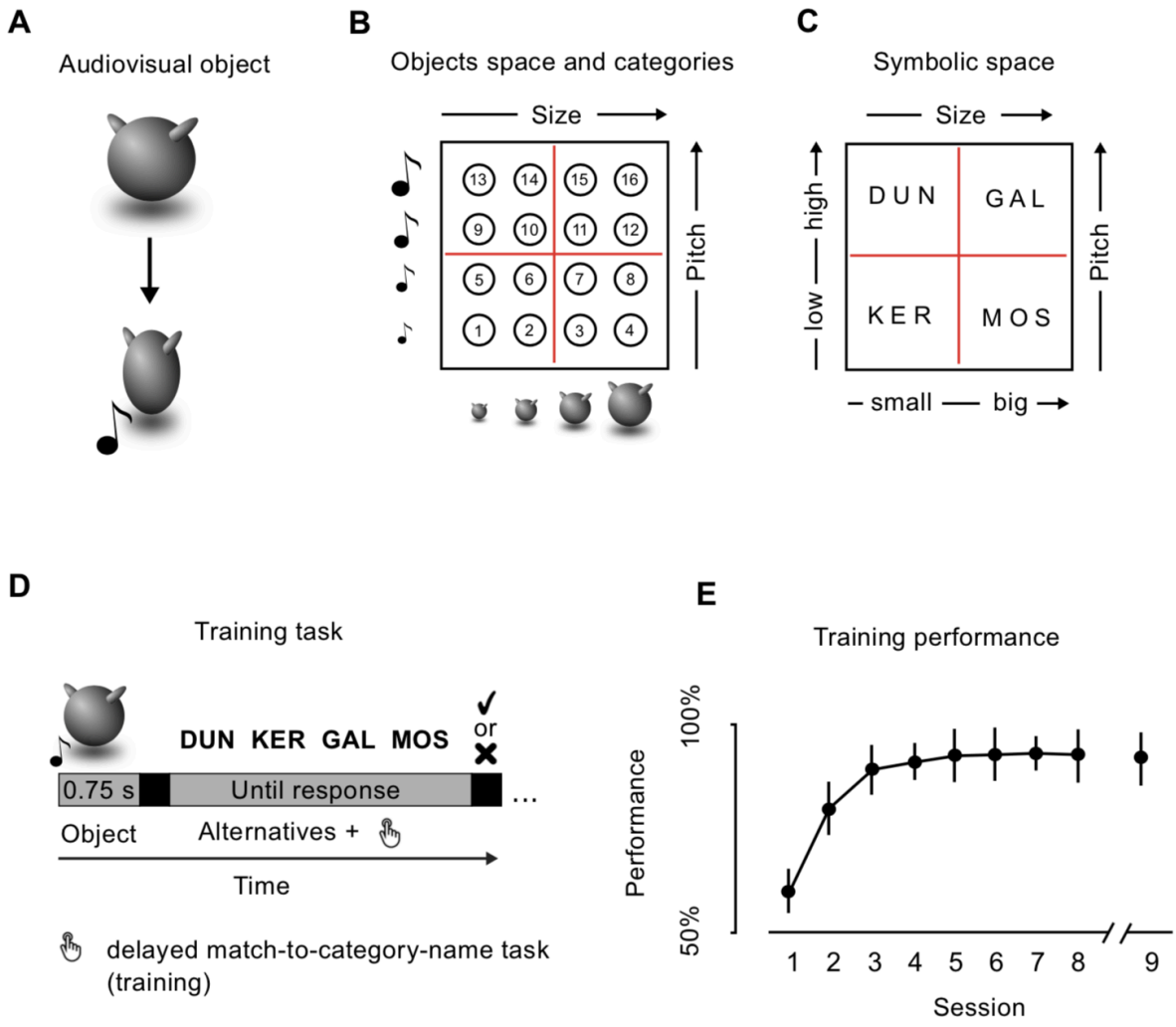

Figure 1.1 - Validation of the behavioural training. A. Example of audiovisual object. **B-C.** 16 multisensory objects are created as unique audiovisual combinations, and they are divided into four categories by means of abstract names. **D.** Participants perform for 9 days a delayed-match-to-category-name task to learning the correct object-name association. **E.** Learning curve shows improvements; performance on session 9 demonstrates a generalization of the categorical rule.

NEURAL MECHANISMS UNDERLYING THE EMERGENCE OF SEMANTIC REPRESENTATIONS

Introduction

A fundamental problem in cognitive science is the “symbol-grounding problem” (Harnad 1980), related to the question of how symbols acquire their meanings. Indeed, a key step for making sense of the rich multisensory world surrounding us is to be able to parse it into meaningful discrete categories, and humans use symbols (such as words or numbers) to construct, recall, and generalise this knowledge. Although even very young children can solve this fundamental act with a striking ease, its neural correlates are still elusive and largely unexplored.

Previous studies investigated how the human brain represents meanings, mostly focusing on the brain responses of adults processing overly known semantic categories, such as that of manipulable objects, food items, animals, or numbers. They report that several regions of the cortex, mostly in the parietal and infero-temporal lobes (see Binder et al. 2009; 2011) contain sufficient information for discriminating concepts both within and across classes, and do so both when they are presented as visual shapes (e.g. Connolly et al. 2012, Clarke & Tyler 2014, Cichy et al. 2014) and as written (e.g. Borghesani et al. 2016; Liuzzi et al. 2019) or spoken (Liuzzi et al. 2015; 2017) words. Crucially, few studies suggested that symbols acquire meaning by linking their neural representations to the ones of the (class of) objects they refer to: this has been suggested to be the case in the case of numbers (Piazza et al. 2007; Eger et al. 2009), color (e.g. Simmons et al. 2007), objects/tools (Chao et al. 1999), and places (Kumar et al. 2017). However, since humans start learning the meaning of words and thus constructing these kinds of representation extremely early in life, it has not been possible to date to witness the neural changes underlying their

emergence. As a consequence, how the human brain solves the symbol-grounding problem remains unknown.

While functional imaging in young children is possible, it is extremely time consuming and difficult to perform. A potentially easier and more powerful way to attack this issue is to engage adult participants in training studies, where they have to learn new concepts by giving names to previously unseen classes of objects or events. Monitoring, through functional neuroimaging, the changes occurring in the neural representations evoked by the stimuli (the symbols and their referents) as a function of learning, should unveil the brain mechanisms underlying the emergence of semantic representations. Behavioural studies already showed that learning to categorise visual objects using labels (that is, linking specific portions of a perceptual space to an abstract symbol and therefore creating a meaning, or a semantic representation, for that portion), alters perception, facilitating categorization itself (Lupyan et al. 2007) and even altering the perception of the objects themselves (Goldstone et al. 1994). This suggests that the way our brain creates new meanings through symbols significantly affects its own internal representation of the external world, and that the brain mechanisms engaged to solve the symbol-grounding problem might be more profound than simply associating two previously separate representations.

In this study, I focused on the neural correlates of learning novel categories of multisensory objects by giving them a name, and asked two specific questions:

i. where and how the new semantic representations emerge, as a function of learning, in the human brain?
ii. how profound are the changes induced by symbolic learning on perceptual representations?

I designed a longitudinal learning experiment where a behavioural training was paired with two fMRI sessions: one before and one after the training period. Participants learned for 9 days to associate novel multisensory objects to written names that represented their categorical identity. I

focused on written words because reading is one of the most distinctive abilities humans developed in the course of the evolution, strongly linked to the act of creating and conveying meanings using symbols, and the processing of which we have a good cognitive and neuroscientific understanding (e.g. McCandliss et al. 2003; Dehaene & Cohen 2011). Also, I opted for using a multisensory object space because previous studies on categorization focused mostly on visual stimuli, thus overlooking other sensory modalities and most of the times not even considering their combinations. In real life, however, we constantly integrate information coming from different sensory inputs to correctly recognize objects (for instance, I recognize an individual by integrating several visual features of her face with the specific sound of her voice), and how this multisensory integration relates with the process of creating semantic representations is ignored.

Before and after learning participants were presented, during an fMRI scanning session, with pseudorandom sequences of the very same set of multisensory objects and visual words. While before learning they performed a simple one-back task on stimulus identity, after the learning period outside the scanner they were actively engaged in an object-name categorization task that explicitly required to associate each word to the correct objects, which is akin to the task they were performing during their training. In this way, I could properly isolate the brain regions involved in the process of grounding symbols to their meaning, and study the effects of this process at the whole brain level.

Methods

Participants. The study included 25 right-handed adult volunteers (fifteen females and ten males; mean age = 22.20, std = 2.74). All participants gave written informed consent, underwent screening to exclude incompatibilities with the MRI scanner, and were reimbursed for their time. The study was approved by the ethics committee of the University of Trento (Italy). Data from 4 subjects were excluded from the analyses given their poor behavioral performance during the second fMRI day (accuracy < 70%). This led to a final sample of 21 participants (thirteen females and eight males; mean age = 21.95, std = 2.58).

Stimulus space. You can refer to Chapter 1, section 1.3 for identical procedures on how the stimulus space was created. For the current Chapter, the relevant figures are Figure 2.1 (A-B-C) and Figure 2.S1.

Stimuli presentation. You can refer to Chapter 1, section 1.3 for identical procedures on how the stimulus space was presented. Behavioural and fMRI sessions were matched for stimuli presentation details.

Experimental sessions. The experiment consisted of three parts: pre-learning fMRI, behavioral training, and post-learning fMRI (Figure 2.1D). During pre-learning fMRI, participants were exposed for the first time to the new multisensory objects and to the abstract names. This allowed recording of the patterns of neural activity evoked by the stimuli when they didn't share any relationship. Starting with the following day, subjects underwent nine sessions of behavioral training outside the scanner. The aim was to teach them the object-name correspondence, an operation requiring parsing of the object space into four categories and connecting each symbol (word) to its meaning (the correct category exemplars). Finally, during the post-learning fMRI, they were again exposed to the same objects and words, now probing their mutual correspondence, and allowing us to record the updated cortical activity. On average, the second fMRI session occurred 9.86 days (std = 1.4 days) after the first one. All the tasks are described below. During both fMRI sessions, and during the first 8 training days, I used 8 out of 16 objects available in each subject' stimulus space (objects: 1 - 3 - 6 - 8 - 9 - 11 - 14 - 16); the remaining 8 were used only during the 9th training session to test for generalization (see below).

Functional localizer. At the start of the pre-learning fMRI session, participants underwent a block-design functional localizer designed to isolate the cortical regions recruited to process visual and acoustic components of our objects, as well as their conjunction. During video mini-blocks, participants were presented with animated objects varying in their size, without any acoustic

component. During audio mini-blocks participants were presented with sounds varying in pitch, without the object visual component. Finally, during multisensory blocks participants were presented with multisensory objects: animated objects varying in size, associated with sounds of different pitch. There were four blocks for each condition (video, audio, multisensory), resulting in a total of twelve mini-blocks of six stimuli each, presented in pseudo-random order. Each block was preceded and followed by 10 s of fixation cross. For each block, participants had to perform a simple 1-back task, pressing a button whenever they detected a repetition of the same stimulus: same size for video blocks; same pitch for audio blocks; same size and same pitch for multisensory blocks.

fMRI tasks. During the first fMRI session, participants performed a simple 1-back task on stimulus identity, where they were presented with the multisensory objects and the four abstract words in pseudorandom order. They were instructed to press a button when they detected an immediate repetition of the very same stimulus (either multisensory object or word). In the case of multisensory objects, they had to take into account both the size of the object and the pitch of the associated sound to provide the correct answer (Figure 2.1E). Each stimulus was presented for either 750 ms (objects) or 500 ms (words), with a variable ISI of 4 +/- 1.5 sec during which a blue fixation cross was presented. There were 4 runs, each one lasting around 7 minutes. Within a run, each stimulus (8 objects and 4 words) was repeated 6 times, resulting in 72 trials per run. There was one target event (1-back repetition) per stimulus, for a total of 12 out of 72 (~17%) expected responses per run. During the second fMRI session, participants were presented with a 1-back task on word-object correspondence, were they had to correctly associate each object to the corresponding name. This task could not be performed before learning given the absence of any categorical knowledge for our stimulus space. Participants were instructed to press the button anytime a multisensory object was followed by the corresponding name (e.g. object 1 followed by the word “KER”), and vice versa (e.g. word “KER” followed by object 1), requiring thus access to newly learned symbolic identity. This resulted in a total of 16 target events (~ 22%) per run. The

number of runs, trials, and stimuli repetition matched the 1-back task on stimulus identity on the first fMRI day.

Behavioral training. You can refer to Chapter 1, section 1.3 for identical training procedures. The relevant figure for the current Chapter is Figure 2.1F.

Neuroimaging acquisition. Data were collected on a 4T Bruker scanner (Bruker BioSpin) with standard head coil, at the Center for Mind/Brain Sciences, University of Trento, Italy. Functional images were acquired using EPI T2*-weighted scans. Acquisition parameters were as follows: TR = 3 s; TE = 21 ms; FA = 81°; FOV = 100 mm; matrix size = 64 x 64; number of slices per volume = 51, acquired in interleaved ascending order; voxel size = 3 x 3 x 2 mm. T1-weighted anatomical images were acquired twice per participant (pre- and post-learning) with an MP-RAGE sequence, with 1 x 1 x 1 mm resolution.

Preprocessing and General Linear Model. Functional images were preprocessed using the Statistical Parametric Toolbox (SPM8) in MATLAB. Preprocessing included the following steps: realignment of each scan to the first of each run; co-registration of functional and session-specific anatomical images; segmentation; normalization to the MNI space. No smoothing was applied. Functional images for each participant individually were analyzed using a general linear model (GLM) separately for the two fMRI sessions. For each run, 22 regressors were included: 13 regressors of interest, corresponding to the onsets of the eight objects, the four words, and the motor response; 6 regressor for head-movements (estimated during motion correction in the pre-processing); 3 regressors of no interest (constant, linear, and quadratic). Baseline periods were modelled implicitly, and regressors were convolved with the standard HRF without derivatives. A high-pass filter with a cutoff of 128s was applied to remove low-frequency drift. I thus obtained one beta map for each stimulus (eight objects and four words) for each run.

Split-half correlation analysis: words and objects identities. First of all I isolated those brain regions representing the identities of the 4 words and the 4 objects. To do that I applied a multivariate approach (see Haxby et al. 2014), implemented in a whole brain searchlight. A sphere with a radius of 3 voxels - selected for consistency with previous studies (Connolly et al. 2012) - was centered in every voxel of the subject- and session-specific datasets. Within each sphere, I conducted a split-half correlation analysis (Haxby et al. 2001) that allowed me to test whether the distributed activity within a brain region differentiates between stimuli. I extracted, within each sphere, the patterns of neural activation across voxels for either the 4 words or the 8 objects, separately. Then I divided the dataset into two halves, and I crossed the neural representations of each stimulus (either words or objects) with each other, resulting in a correlation matrix with 4x4 entries for words, and 8x8 entries for objects (Figure 2A-D): here, the correlation between matching stimuli coming from the two different halves laid on-diagonal, while the correlation between non matching stimuli laid off-diagonal. If the activity in the ROI is differentiating between stimuli identities (that is, is representing differently the four words or the eight multisensory objects), the mean difference between Fisher-transformed values on-diagonal versus off-diagonal, resulting from all the possible combinations of the two dataset halves, should be positive. For each sphere, the resulting correlation score was stored in the center voxel, therefore I obtained one correlation map per subject, per session, and per type of stimuli (words or objects). Single-subjects' correlation maps were then submitted to group-level analysis to reveal significant clusters of voxels where multivariate information was sufficient to distinguish different words and different object identities. In the specific case of object identities, to be sure that the resulting clusters were sensitive to multisensory information and not to one of the two sensory features (that is, differentiating objects only basing on their size or on their pitch), I additionally run two corresponding searchlights but looking for brain regions responding to unimodal variations between objects. I used the union of these two resulting maps as exclusive mask for the group-level analysis on object identities, therefore guaranteeing that the resulting clusters were sensitive to multisensory conjunction only, that is the real definitional criteria of our individual object identities. Spherical ROIs with radius of 8mm were created around the peak voxels, to be sure that

following analyses were conducted on Regions of Interest (ROIs) of matching voxel size. Corresponding results were obtained when the entire clusters were considered.

Crossmodal correlation analysis. In the brain regions individuated from the two split-half correlation analyses, encoding the identity of either the four words or of the 8 objects, I run a crossmodal correlation analysis, where I attacked directly the core question of the study, that is the rise of mutual correspondence between the representations of objects and the names (symbols) used to denote them. By looking in these two areas I put to direct test the hypothesis that brain regions representing either words or objects change their activity to reflect the acquired association with the corresponding referent or symbol, respectively. In the crossmodal correlation analysis I divided the dataset into categories of objects (e.g. objects 1 and 3, or objects 6 and 8) and words, and I crossed the neural representation of each object category to the neural representations of each name (e.g. 'KER' or 'GAL'), resulting in a 4x4 correlation matrix. I reasoned that if voxels within an ROI represent the correct category-name association (e.g. the category composed by objects 1 and 3, and the word 'KER'), then the correlation of neural patterns corresponding to matching stimuli (on-diagonal) should be higher than that of non-matching ones (off-diagonal). Thus, the mean difference between Fisher-r-to-z-transformed on-diagonal vs. off-diagonal values is stored for each subject, as summary of the information for the ROI, and subjects' correlation scores are later tested against a null hypothesis of no correlation at the group level. Additionally, to avoid overlooking other potential brain regions that could contribute to this association, I implemented the same ROI analysis in a whole brain searchlight (for parameters, see above Split half correlation analysis).

Decoding stimulus modality (words vs. objects). To investigate the contribution of the brain regions individuated by previous analyses, I wanted to quantify, in each region, the extent of abstraction in the representation of semantic classes. The neural signature of full abstraction from stimulus modality would correspond to an absence of residual information relative to it, that is the impossibility to decode whether, at any given trial, subjects were presented with a word or the

given corresponding object. In order to test this I implemented a decoding approach, because the higher the information on the stimulus modality, the higher the performance of a classifier trained with that that information to predict the incoming modality of an independent stimulus. I used a leave-one-run-out scheme to train and test a Linear Discriminant Analysis (LDA) in correctly predict the modality of the incoming stimulus, and I stored each subjects' and ROI's accuracy for later group-level test against a null hypothesis of chance performance (50%). Corresponding results are obtained using a Support Vector Machine (SVM).

Perceptual learning and sensory segregation. To investigate is the changes induced by learning could be traced down to the activity of sensory regions, I focused on the representational geometries of objects in those brain regions that responded separately for their size or their pitch. ROIs responding to the visual and to the acoustic components of our multisensory objects were isolated on pre-learning imaging data. I selected brain activity evoked at group-level ($p<.001$; FWE corr.) by objects presentation during the 1-back task on stimulus identity. I masked the signal with the group-level results ($p<.001$; FWE corr.) of the functional localizer for either the visual and the acoustic modality. This resulted in a bilateral network wherein the Lateral Occipital Complex (LOC) and the anterior portions of the Superior Temporal Gyrus (STG) responded to the visual and to the acoustic components of our stimuli, respectively (Figure 2.4A). All clusters were binarized and used as regions of interest in the following analyses. In absence of a priori hypotheses on the lateralization of sensory signals, bilateral ROIs were used. In order to investigate whether the activity in sensory areas responding to visual and acoustic components of our objects changed after learning, I extracted neural dissimilarity (1 - Pearson's correlation) between pairs of all objects varying along one sensory dimension only (e.g. between object 1 and object 3, that varied in their size but had the same associated sound, Figure 2.4B), and considered their difference between the two fMRI sessions as dependent variable. I conducted a 2x2 repeated measures ANOVA looking for the interaction between the two ROIs (LOC and STG) and the two sensory modalities (distance between objects with different size but same sound, and vice versa). This approach was motivated by the fact that I wanted to describe whether the act of connecting objects

to their names (which is a fundamental step in the symbol-grounding problem) affected the perceptual representations of the referents (here the objects). By taking advantage of a longitudinal neuroimaging study, I could compare their neural representations after learning, during a semantic categorization task, with a pre-learning condition where no semantic information could be retrieved, not even automatically, because it was not part of participants' knowledge.

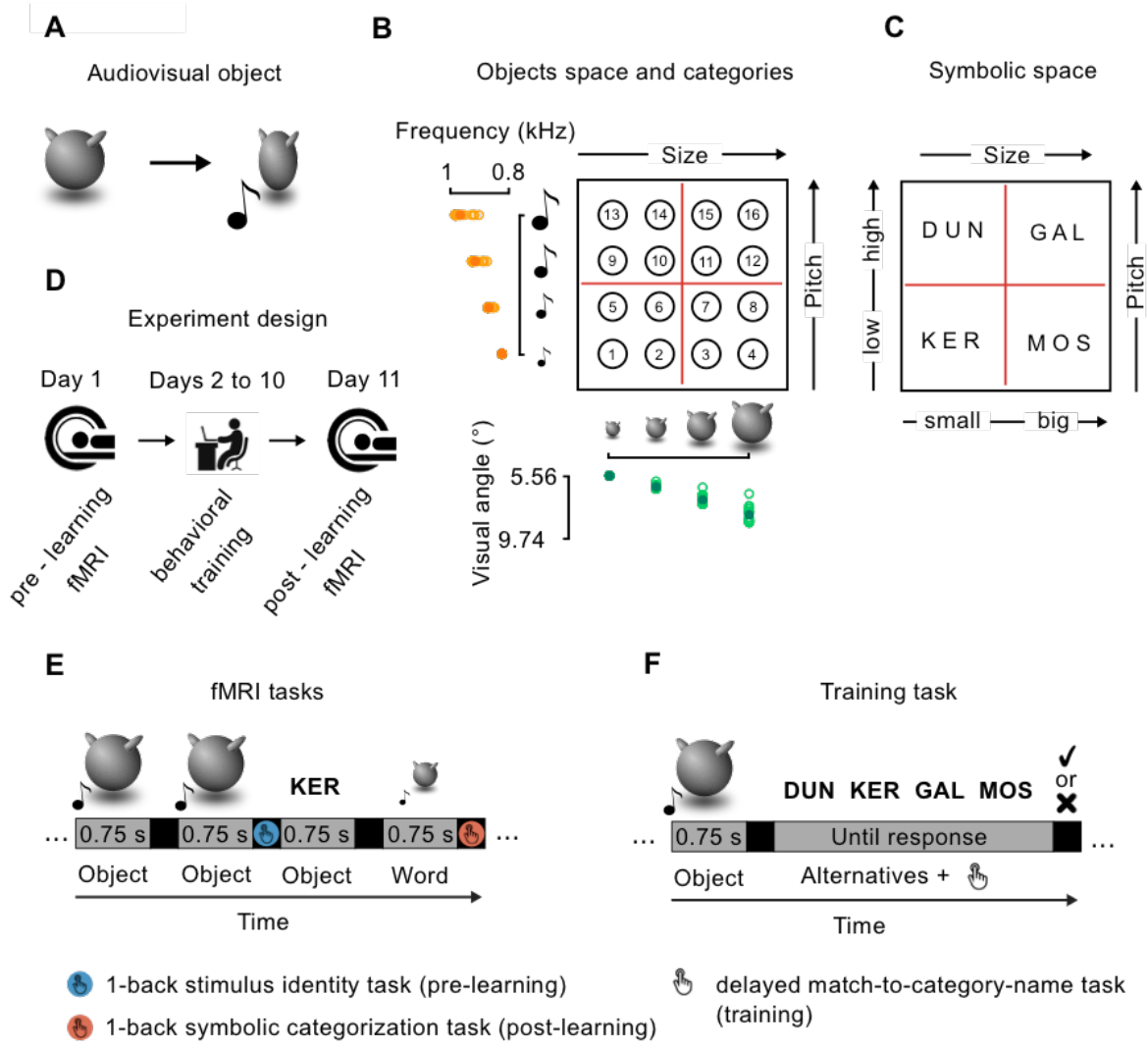

Figure 2.1 - Methods. A. Example of audiovisual object. **B-C.** 16 multisensory objects are created as unique audiovisual combinations, and they are divided into four categories by means of abstract names. **D.** Experimental design: two fMRI sessions (pre and post learning) are paired with a 9-days long training. **E.** Tasks performed in the fMRI in the two days. **F.** Training task.

Results

During 9 daily sessions of training, 21 healthy volunteers learned to recognise 8 new audiovisual animated objects by associating them to one of 4 novel words (Figure 2.1A-B-C). Objects were created by varying parametrically and crossing orthogonally the size of an abstract shape and the pitch of the sound produced during a little animation (see Methods). The unique object identity was therefore multisensory in nature and the correct names could be recovered only by attending to the specific combinations of the two sensory features, thus through multisensory integration. I paired this behavioural training with two fMRI recording sessions: one pre- and one post- training (Figure 2.1D, see below). During both sessions, participants were presented with pseudo-random sequences of the same 8 audiovisual animated objects and the 4 written words corresponding to their category names. Before learning, subjects performed a 1-back task on stimulus identity, while after learning they performed 4 runs of a 1-back task on word-object identity explicitly probing the newly acquired object-name associations (Figure 2.1E, see Methods)

Behavioral training results. The behavioural training consisted of 9 daily sessions outside the scanner, lasting ~ 20 min each (see Methods). During each training session, participants were first presented with one of 8 randomly selected audiovisual object per trial (training set), and then with the 4 written words in shuffled order (Figure 2.1F). They were told that each object type belong to a specific category which had a specific name, and that they had to select the correct word, receiving feedback on their performance on every trial. During the last training session, and without being notified, participants were also presented with 8 novel stimuli. These consisted in specific combination of size and pitch that were absent in the training set, allowing testing for generalisation (Supplementary Figure 2A). The learning trajectory indicates an increment in performance from session 1 to 4, while from session 5 on participants maintained unchanged their accuracy level (session 1: 61 ± 18%; session 2: 76 ± 17%; session 3: 83 ± 13%; session 4: 87 ± 11%; session 5: 89 ± 10%; session 6: 90 ± 11%; session 7: 91 ± 8%; session 8: 91 ± 7%), indicating a period of consolidation (Supplementary Figure 2.2B-C). Performance collected on

session 9 confirmed both the successful learning of the name-objects association and its generalization. Performance was high for both sets (training set: mean accuracy = 88.12%, std = 5.93%, different from chance t = 29.44, p < .001; generalization set: mean accuracy = 75.03%, std = 10.68%, difference from chance, t_{20} = 10.74, p < .001), even though it was lower for the generalization set (t = 5.06, p<.001), indicating that while generalisation did occur, it was not perfect (Supplementary Figure 2.2D-E).

Neuroimaging results

The emergence of semantic representations in VWFA. I started by isolating those brain regions where multivariate activity differentiated between words' identities. In a whole brain searchlight I implemented a split-half correlation analysis (see Methods) that allows one to test whether a brain region represents the different identities of the stimuli considered. Before learning, thus when words did not correspond to any object class (namely, they had no meaning), this resulted in two significant clusters: one in the right lingual gyrus ($MNI_{x,y,z}$: 15, -76, -2; t = 5.27) and one in the left inferior fusiform gyrus ($MNI_{x,y,z}$: -45, -61, -8; t = 4.85), a region known as Visual Word Form Area (Dehaene & Cohen 2011) (Figure 2.2A-D). Then I asked whether the neural representations of those stimuli were modified during learning, and more precisely whether their response to the words became more similar to the ones evoked by the corresponding objects (e.g. the word KER and object 1) then to the non corresponding ones (e.g. the word KER and object 16). This classical view of the role of these regions in reading (representing the lexical orthographic pre-semantic stage of processing) would predict no trace of semantic coding. I applied, in these two ROIs, a crossmodal correlation analysis (see Methods) that allows one to reveal whether words and the corresponding objects are represented similarly, thus suggesting a shared neural code between the symbols and their specific referents. I found a significant result in the VWFA (t = 3.74, p = .001) but not in the lingual gyrus (t = 0.62, p = .54)(Figure 2.2B). This word-object correspondence was not present before learning (t = 1.6, p = .13), when no knowledge of the object-name association was present.

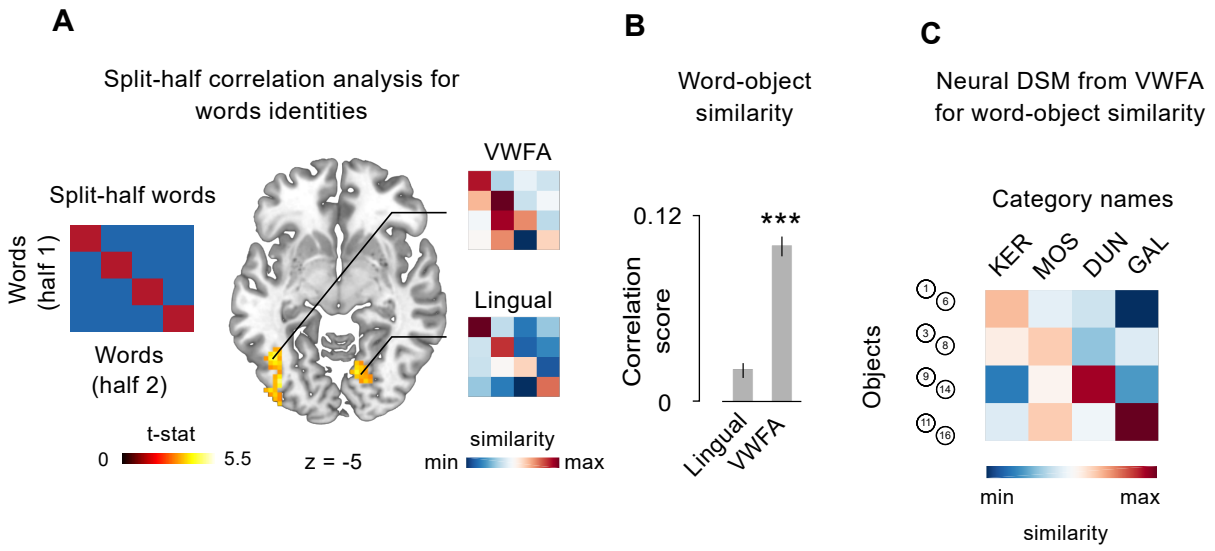

Figure 2.2 - A shared representation for objects and their categorical names in VWFA. A. Split-half correlation analysis reveals individual word representations in VWFA and lingual gyrus. Neural DSM for illustrative purposes. **B-C.** After learning, the representations of words in VWFA become more similar to the ones of the corresponding objects.

The emergence of specific representations for each multisensory trained objects. Next, I applied the same procedure looking for brain regions showing specific pattern of responses to the different object identities. While a whole-brain searchlight on the brain activity before training revealed that the objects were not discriminable, after the training three brain regions locally contained distributed activity sufficient for discriminating across all individual objects: the left angular gyrus ($MNI_{x,y,z}$: -36, -67, 26; t = 6.8), the left middle frontal gyrus (MFG)($MNI_{x,y,z}$:-33, 47, 18; t = 7.82), and the right Inferior Frontal Gyrus (IFG)($MNI_{x,y,z}$: 51, 35, -6; t = 6.98)(Figure 2.2D). In these regions the signal coming from unisensory areas is integrated to give rise to the representation of single multisensory objects.

The emergence of semantic representations in L-AG. Then, within these three ROIs I applied the crossmodal correlation analysis to verify whether concurrently with the differentiation between objects they also developed a response of category that was similar across words and objects (the same analysis performed in the VWFA). Of the three regions, only the left angular gyrus (t = 3.53, p = .002) displayed this coding feature (MFG: t = 0.9, p = .36; IFG: t = 1.73, p = .10)(Figure

2.2E). This pattern of similarity in the left Angular Gyrus was absent before learning (t = 1.04, p = .31).

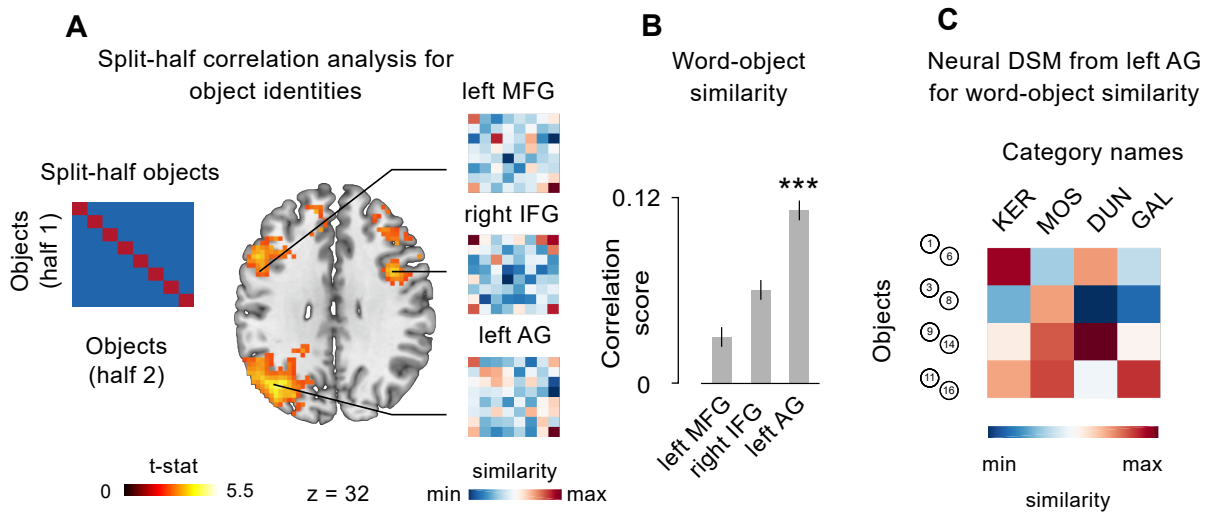

Figure 2.3 - A shared representation for objects and their categorical names in left AG. A. Split-half correlation analysis reveals individual object representations in left AG, left MFG, right IFG.Neural DSM for illustrative purposes. **B-C.** Object representations in the left AG are more similar to the ones of the corresponding names.

A distributed network encoding object-name association. Although our previous results already led to drive interesting conclusions on the development of corresponding associations between symbols and their referents, I further investigated whether any other brain region represented this correspondence. I implemented the crossmodal analysis in a whole brain searchlight that revealed an additional set of brain regions that, after learning, represented the association between objects and their categorical names: other than the left fusiform gyrus/VWFA ($MNI_{x,y,z}$: -51, -61, -12; t = 6.7) and the left Angular Gyrus, here also extending to the Superior Parietal Lobule ($MNI_{x,y,z}$: -27, -58, 46; t = 6.6), that confirmed the findings revealed in the previous ROI approach, categorical and modality invariant information was also present in the right Superior Parietal Lobule ($MNI_{x,y,z}$: 21, -73, 54; t = 6.6), in the right Inferior Parietal Lobule ($MNI_{x,y,z}$: 57, -46, 46; t = 5.1), in the right Angular Gyrus/ Temporo-parietal Junction ($MNI_{x,y,z}$: 54, -67, 34; t = 5.25), and in the right Hippocampus ($MNI_{x,y,z}$: 27, -22, -6; t = 5.14) (Figure 2.4A). Before learning, no region showed this effect.

Tolerance to variations in the stimulus presentation modality. How do these areas support the emergence of object-name association? While VWFA and the left Angular Gyrus also encode the identity of single word and of single objects, respectively, and developed a similarity between matching symbol-referent pairs, the right-lateralized ROIs emerged only after the crossmodal searchlight. This raises the possibility that these brain regions show a higher level of invariance to the presentation modality, which could be potentially useful to construct the object-name association. To test this hypothesis I implemented a decoding procedure where I trained a Linear Discriminant Analysis (LDA) classifier to predict the sensory modality of the presented stimulus (word or object) (see Methods). Although I selected these ROIs for the fact that they similarly respond to objects and their corresponding names, I don't know the degree of their sensitivity to differences in the two presentation modalities - which, however, should be trivial for a classifier to capture, given the extremely different lower-level perceptual features between the audiovisual objects compared to the written words. The classifier accuracy was indeed very high, especially in the left-lateralized ROIs (VWFA: mean accuracy = 64%; $t = 6.14$, $p = 5.25 \times 10^{-6}$; L-AG/SPL: mean accuracy = 59%; $t = 5.15$, $p = 4.77 \times 10^{-5}$), and decreased as I moved to right regions, and from superior cortical structures down to the hippocampus, where the performance of the classifier did not diverge from chance (R-SPL: mean accuracy = 59%; $t = 4.70$, $p = 1.35 \times 10^{-4}$; R-IPL: mean accuracy = 56%; $t = 3.66$, $p = .001$; R-AG/TPJ: mean accuracy = 55%; $t = 2.41$, $p = .026$; R-HPC: mean accuracy = 52%; $t = 1.60$, $p = .12$)(Figure 2.4B).

Correlation with behavioural performance. Finally, I correlated the crossmodal correlation scores in the areas of our network with the behavioural performance collected at the end of the training, before the second fMRI session. I found that the degree of similarity between the neural representations of the objects and their corresponding names in the right hippocampus significantly correlated with overall behavioural performance during the last day of training ($R = 0.58$, $p = .006$)(Figure 2.3D), while none of the other areas showed this effect (all $p > .17$). Interestingly, the correlation remained significant when restricting the categorization performance during the last training day to only the subset of novel (generalized) objects ($R = 0.54$, $p = 0.012$).

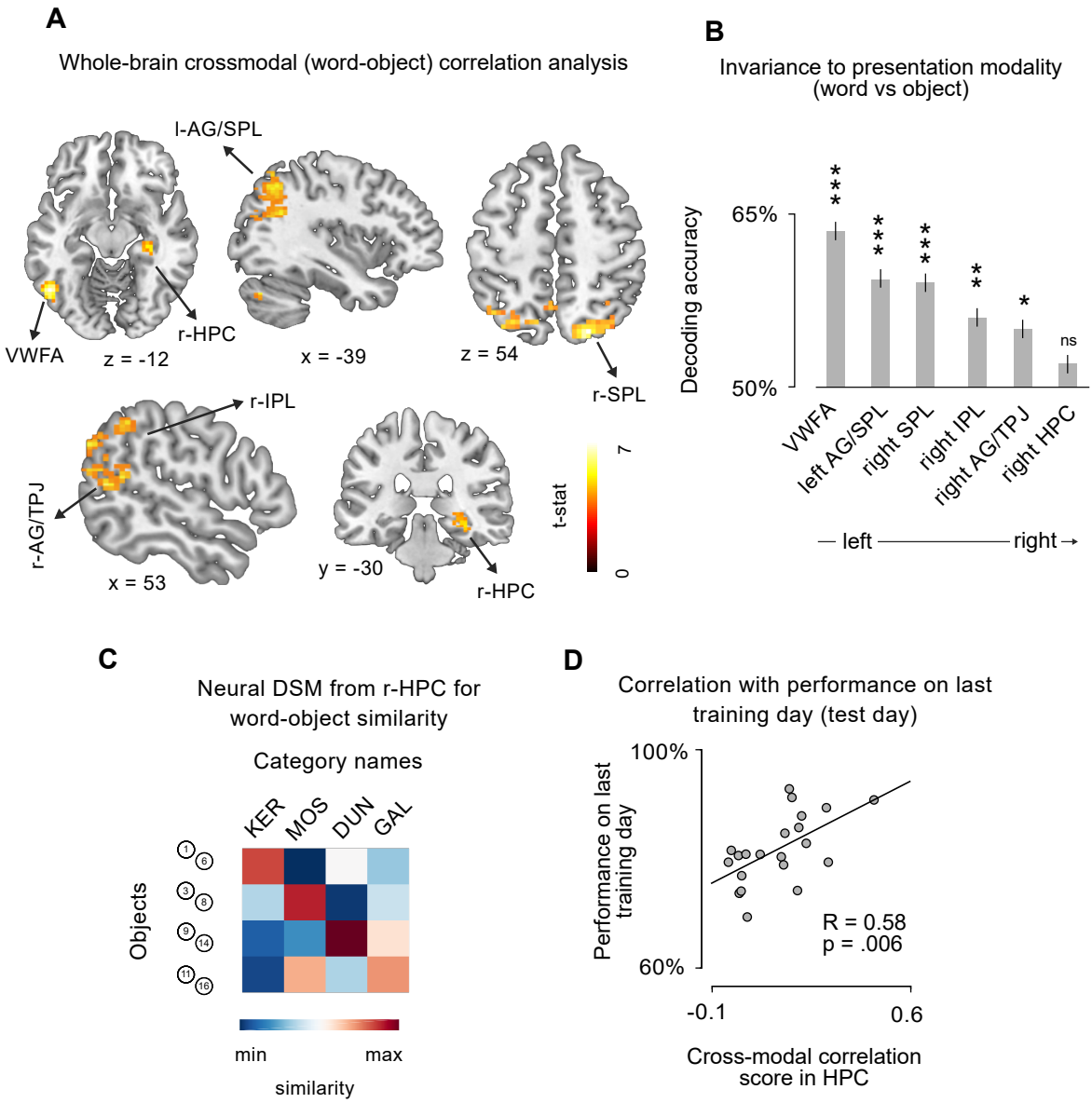

Figure 2.4 - A temporo-parietal network supporting the emergence of semantic representations and the role of the right Hippocampus. A. Whole-brain results of the crossmodal correlation searchlight (S/IPL = Superior/Inferior Parietal Lobule; HPC = Hippocampus). **B.** A LDA classifier trained to discriminate the presentation modality shows decreasing performance as we move from left to right ROIs, and specifically down to the right HPC. **C.** Neural DSM of crossmodal correlation in r-HPC. **D.** Crossmodal correlation score in r-HPC significantly correlates with behaviour during the last training day (test day), when novel and generalized objects are presented.

Perceptual learning and sensory segregation. Finally, I focused on the second question I wanted to attack with this experiment, that is to what extent the effects of symbolic categorical learning also affect sensory coding of the object features. With our split-half object-based correlation analysis I

already demonstrated an increased sharpening of the representation of individual multisensory object identities after learning, during our symbolic categorization task, in a frontoparietal network. Does this effect extend to lower level perceptual regions, that could potentially inform and support higher-level multisensory integration, necessary to the task at hand? To answer the question, I looked for evidence of sharpening of the differences along unisensory modalities in their specific sensory regions (e.g. differences in sound frequency within acoustic sensory cortices) and, additionally, in the sensory regions responding to the opposite modality (e.g. differences in sound frequency within visual cortices responding to size). I approached the problem by first selecting, through a independent functional localizer, visual and acoustic ROIs - bilateral Occipital Cortex (LOC) and anterior Superior Temporal Gyrus (STG) respectively - responding selectively to the visual and acoustic components of our multisensory objects (see Methods)(Figure 2.5A). Thus, on the basis of previous report of plastic properties of sensory areas showing an increase in precision of visual and acoustic information as a function of unisensory learning (e.g. Jiang et al. 2007; 2018), I asked whether similar perceptual learning effects can be detected with our design. To answer the question, I investigated if and how unisensory regions modify their representations of each sensory feature separately. In each of the two sensory ROIs, I compared the neural representational dissimilarity (1 - Pearson's r)(see Methods) before and after training, between objects varying along each of the two dimensions separately, and used an ANOVA to compare perceptual learning (indexed by the difference in dissimilarity between the two fMRI sessions) across regions and sensory features (Figure 2.5B). The ANOVA revealed a significant interaction between changes along sensory dimensions (size vs pitch) and ROIs (LOC vs aSTG)($F = 8.13$, $p = .009$)(Figure 2.5C). This indicated that training induced a form of sensory segregation in sensory regions: both regions developed an increased sensitivity for the preferred sensory dimension, and that was paired with a decreased discriminability between differences along the non-preferred one.

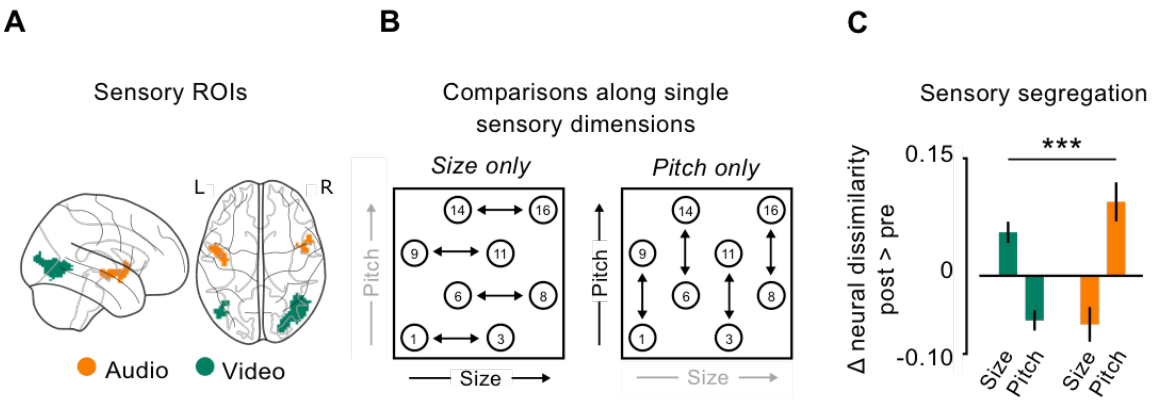

Figure 2.5 - Perceptual segregation in sensory areas. A. LOC and STG respond to the visual and acoustic components of our objects, respectively. **B.** Neural dissimilarity is compared between objects varying only along one sensory modality, and between the two session (pre post) and ROIs. **C.** Significant interaction reveals sensory segregation.

Discussion

In this study, I attacked the question of how novel semantic representations, that are meaningful representations of object classes conveyed by symbols, emerge in the brain. I reported three main findings: i) an increased similarity between the neural representations of symbols (words) and their referents (the objects they refer to) emerge in the VWFA and in the left Angular Gyrus, two areas that also encode the identity of the individual words and individual objects, respectively; ii) beyond these left-lateralized regions, also a set of right-lateralized areas - encompassing superior and inferior parietal cortices and the hippocampus - support the emergence of this shared neural similarity and, among those, the right hippocampus shows full invariance to presentation modality and strong correlation between its multivariate activity and behavioural performance on the last training day (test day); iii) processing the object-name correspondence also drives changes in the representation of multisensory objects, in the form of increased sensory segregation in those lower perceptual regions that code for either their visual or their acoustic component.

The posterior portion of the left occipitotemporal sulcus plays a key role in reading (Dehaene & Cohen 2011), being sensitive to the presentation of written words over objects (Cohen et al. 2000), and invariant to changes in the location (Cohen et al. 2000), in case (Dehaene et al. 2001, 2004), or font (Quiao et al. 2010) of the presented words. The area was named Visual Word Form Area (VWFA), and its role as inferred from imaging studies was consistent by neuropsychological evidence: patients with lesions to the VWFA develop severe pure alexia, that is the loss of the ability to efficiently identify visually presented words irrespective of their lexical or semantic status (Cohen et al. 2000, 2003; Starrfelt et al. 2009; Mani et al. 2008). These results have been replicated many times (e.g. Jobard et al. 2003 for a meta-analysis), and have been further strengthened by longitudinal training studies showing that the degree of literacy of both adults and children correlates with the activation of this area during word identification (Dehaene et al. 2010; Cantlon et al. 2011). In the present study, I proved that the VWFA differently represents, in its multivariate activity pattern, the identity of short pseudo-words that subjects did not encounter before and thus had no meaning. This is the first time, to our knowledge, that individual word identities could be differentiated from the activity patterns in VWFA, and this is likely due to the presence of only 4 words in our design. Moreover, and crucially, I demonstrated that after a symbolic categorical training where participants learned to map these words onto categories of novel multisensory objects, the representational geometry of the VWFA also reflected the newly acquired object-name similarity. This result significantly extends the previous body of work linking the VWFA to an orthographic level of encoding only, by showing that its plastic properties might also support the link between the visual form of a word and the meaning it refers to. Recent works suggested indeed that the multivariate activity of this area is affected by the semantic content of the words if that is relevant for the task that subjects are performing. Want et al. (2018) report that the multivariate activity patterns of the VWFA evoked by known words during different semantic tasks significantly correlates with the semantic information they convey and that is relevant for the ongoing task: the similarity of the neural representation of words in the VWFA between two words like “doctor” and “teacher”, or between “hospital” and “school” is high during a taxonomic task that

requires to judge the similarity of words on the bases of their taxonomic membership (people vs locations), while it is low during a thematic task that enhance contextual information, and that would predict an higher similarity between, for instance, “doctor” and “hospital”, or “teacher” and “school”. Crucially, this pattern was inverted when participants were actively engaged in a thematic semantic task, showing that the multivariate activity of the VWFA represents semantic information and adapt to the ongoing task-setting and behaviour. In my experiment I could demonstrate that the plastic properties of this area allow for the emergence of a shared neural code between the words and their referent that was not present before learning.

The emergence of object identities and semantic representation in the left Angular Gyrus

While my previous findings showed that a brain region coding for lexical/orthographic properties of novel words (VWFA) developed, after learning, an increased similarity between these words (the symbols) and the objects belonging to the category they refer to (their referent, or meaning), I additionally seek for the specular pattern by looking for those brain regions where object identities were represented, to see whether they also developed a response to words that reflected the newly acquired object-name similarity. Surprisingly I did not find evidence of individual object representations before learning, despite an almost perfect task performance (>90% of correct responses). A possible explanation for this null finding is that participants were not integrating the visual and acoustic features in a single individual combination, but they were rather processing size and pitch separately to solve the one-back identity task. Interestingly, such individual representations emerged after learning, when participants had to recover their names, in a fronto-parietal network encompassing the left Angular Gyrus, the left middle frontal gyrus, and the right inferior frontal gyrus. While a central role of the L-AG in supporting semantic memory is widely accepted (Binder et al. 2009; Binder & Desai 2011), little was previously known with respect to its precise function and nature of semantic coding schemes. Given its anatomical position, the L-AG well suits as convergence zone (Damasio 1989, Binder & Desai 2011) to integrate information coming from lower sensory regions. Indeed, recent studies indicated its causal role during

multisensory integration (Bonnici et al. 2016; Yazar et al. 2017). These findings are coherent with our discovery that the representation of individual objects which identity is defined by the integration of specific visual and audio properties emerge in this region. Crucially, the emergence of a representation of the individual objects was paired with an increased representational similarity between the objects and the symbols that identify them. This is suggestive of similarities with the case of numbers, that in the IPS, known to represent quantities, evoke similar representations both when quantities are presented in their symbolic (as Arab digits) and non-symbolic (as dots) form (Piazza et al. 2007). This might indicate that the human brain employes a parsimonious solution to the symbol-grounding problem, mapping symbols representations directly onto those neural circuits that respond to the objects these symbols refer to. Less clear, in this respect, is the role of the frontal regions that, although representing individual object identities, did not develop similarity between their representation and the one of their names. Previous studies in the field of object recognition and categorization indicates that lateral prefrontal cortices are involved in the process of recognizing objects and their categories (e.g. Riesenhuber & Poggio 2002, Jiang et al. 2007; 2018) and that these effects are modulated by the task (Roy et al. 2010; Van der Linden 2014). Also, it has been shown that the same areas contribute to support working memory in a variety of perceptual and categorization tasks (Lara & Wallis 2015; Miller et al. 2018). Although with the current experiment I can not find a precise answer to what contribution these regions are actually offering to the symbol-grounding problem, I could speculate that their role might be more related to the act of holding in memory the identity of the current object which, being ambiguous and more difficult to discern compared to words, might require an extra effort, for which their contribution could be essential.

The crossmodal network and the role of the right hippocampus

A crucial step in our analyses was to search, using a whole brain searchlight, for other regions where the object-name similarity was present after learning. This was motivated by the fact that to solve the symbol-grounding problem the association might emerge not only in the specific regions

representing the symbol or its referent(s), as I showed above, but also in separate areas, less dependent to the perceptual format of the incoming stimulus, but coding for their similarity in a more abstract way. I did find a set of regions responding to this criterion in the right parietal and in the hippocampal cortices. Crucially, I saw a pattern of increasing abstraction (as indicated by a decrease in stimulus presentation format classification accuracy) as I moved from the VWFA and the left AG to the right hemisphere and down to the right hippocampus, where the performance of our classifier was the worst. A possible explanation of this null result in the HPC is the well established problem of signal loss from the medial temporal lobe (Schmidt et al. 2005; Bellgowan et al. 2006; Olman et al. 2009), that however wouldn't fit with the strong positive result of the cross-modal correlation analysis. A more interesting and more plausible alternative explanation posits that the hippocampus might be truly involved in constructing the relevant semantic association between a symbol and its referent, thus acting as interface to support the symbol-referent association. Indeed, previous studies showed that the hippocampus is crucial in representing the association between the neural representations of objects or, separately, words pairs (Spiers et al. 2001; Giovanello & Keane 2003; Clark et al. 2018). Additionally, it contains neurons that are highly selective to specific stimuli identities (e.g. pictures of a specific place) but highly invariant to the presentation modality of its identity, responding to the same stimulus identity whether it is presented as a picture, as a hand drawing, or even as a spoken or written word (Quiroga et al. 2005; 2012). I observed a strong correlation between the degree of specificity of the similarity between words and their relative objects in the right hippocampus and the performance on the last training day, during which participants were precisely tested in matching the objects to their names. Crucially, in this last testing session I presented participants with the same old objects that they have been trained on and also with novel exemplars, representing audiovisual combinations that subjects had never seen before. The significant correlation between HPC object-name similarity with the behaviour remained also when I considered only the new, generalized, objects. This suggests that the hippocampus likely holds a representation that goes beyond a simple episodic association of a particular object exemplar with a particular word. A recent study by Blumenthal et al. (2017) reported the case of an amnesic patient with well documented

hippocampal lesion, who showed severe deficits in producing semantic features for well-known object concepts when they referred to contextual informations, such as how to typically use an object or where to find it. This report suggested that semantic and episodic memory might not be completely separated as classically assumed. Recent theoretical works (Mack et al. 2017; Morton et al. 2017) suggests that the hippocampus might actually contribute in the course of concept learning by means of pattern separation and pattern completion, that allow to differentiate overlapping experiences for behavioural purposes, as well as integrating those aspects of these experiences that shares commonalities. This might provide a framework to interpret the role of hippocampus in our experiment: during and after the course of the training, it might have developed a representation of object categories by integrating all the common aspects of individual instances of a category and by separating, or abstracting from, those details that do not repeat themselves across the different expositions of the same category. In the case of different exemplars of KER, for instance, the hippocampus might have developed a representation of the core definitional aspects of the category KER, that of being a small object producing a low sound, without representing the perceptual variability within the category. This would allow for a later generalization to novel exemplars when their linguistic label had to be retrieved - exactly as we observed. In this perspective, the hippocampus might play a crucial role in solving the symbol-grounding problem. Future studies should further investigate this idea with complementary methods to non-invasive neuroimaging. One possibility is the study of clinical patients with lesions to their hippocampus, that might reveal that they are indeed unable, or less proficient, to generalize an associative rule such as the one existing between an object and its name to novel exemplars. Additionally, intracranial recordings on epileptic patients that are implanted for clinical reasons should address whether an increased neuronal response encoding the association between an object and its name emerge during and/or after learning, as it happens for object-object associations (Ison et al. 2014).

What remains surprising is the right lateralization of the peak I observed. The right hippocampus is usually associated to visuo-spatial memory, which doesn't bear any relevance to our experimental design, unless we interpret our novel semantic space as a cognitive space (Bellmund et al. 2018)

which structure can be captured in a spatial format. This is not a novel idea, and an increase number of studies is showing how brain regions holding specific spatially-tuned representational code recruits the same mechanisms to represent abstract information (e.g. Constantinescu et al. 2016, Garvert et al. 2017) in the form of an internal cognitive map of memories and concepts (Behrens et al. 2018; Bellmund et al. 2018). A very recent study by Theves et al. (2019) indeed found that the hippocampus encodes distances between well-known objects that are learned, by adult participants, as specific points of a novel bidimensional space. Future studies should more directly address this possibility and explore whether and how these mechanisms are crucial during semantic learning.

Perceptual segregation in sensory regions

Finally, I investigated the effects of symbolic categorization on the neural representations of objects in perceptual regions. While the acoustic and the visual components of the audio-visual objects were conjointly encoded in the L-AG after training, the two sensory features defining each object identity were also separately processed in perceptual areas of the visual and acoustic pathways (Mishkin et al. 1983; Rauschecker & Tian 2000; Obseler et al. 2008). However, learning to map object identities onto the representation on words (and viceversa) changed anterior STG and LOC responses to sensory stimuli, which developed an enhanced sensitivity to differences along their preferred sensory modality: i.e., acoustic areas became more sensitive to differences in the sounds produced by objects; visual areas becomes more sensitive to differences in the size of the objects. Concurrently, they developed a decreased sensitivity to differences along the non-preferred one: i.e., acoustic areas became less sensitive to differences in the size of the objects; visual areas became less sensitive to difference in the sound produced by the objects. I relate these results to two well-known behavioural effects occurring during categorical learning (Gibson & Gibson 1955; Gibson & Walk 1956; Vanderplas et al. 1964). In tasks were subjects learn to discriminate objects on the basis of single dimensions, while their sensitivity to that dimension increases (so called ’acquired distinctiveness’ effect (Lawrence 1949), their sensitivity to

concurrent changes in task irrelevant dimensions decreases (so called 'acquired equivalence' effect (Waller 1970)). These observations lead to the idea that training 'warps' representational spaces such that the perceptual distance (or dissimilarity) between features changes (Harnad 1987; Goldstone 1994a,b). Past imaging work supported this idea but were restricted to unisensory visual (e.g. Kourtzi et al. 2005; Op de Beeck et al. 2006; Op de Beeck & Backer 2010; Folstein et al. 2013; Brants et al. 2016) or acoustic stimuli (Ahveninee et al. 2011, Ley 2012, 2014; Bao 2015). Whether and how these effects extend to multisensory stimulations has been largely ignored. Lemus et al. (2010) for example recorded single neurons in macaque monkeys discriminating, on interleaved trials, between two tactile or two acoustic stimuli. While several neurons in the somatosensory cortices and primary auditory cortex responded to both visual and auditory stimuli, the stimulus identity could only be decoded from responses to their principal sensory modality. Thus, the authors suggested that during multisensory stimulation the representations of the different sensory modalities compete against each other, and sensory cortices select one over the other, according to their perceptual preference. The results of the present experiment are congruent with this view, as I observed that within sensory regions, information along the non-relevant sensory modality was reduced/suppressed (acquired equivalence), in favour of higher sensitivity for relevant sensory differences (acquired distinctiveness). These kinds of "suppressive" effects may be entirely overlooked in multisensory stimulation experiments where the different sensory features of stimuli often do not orthogonally vary but, rather, are correlated and thus predictive of one another. In this experiment, on the contrary, the two sensory features varied orthogonally, such that allowing one modality to interfere with the encoding of the other would reduce accuracy in stimulus recognition, the task that subjects were asked to perform. In this sense, the amount and type of multisensory integration that can be observed in early sensory cortices might be crucially determined by the task and stimuli features used during the experiment at hand, and future work should further and directly investigate their specific role in influencing multisensory information coding in sensory areas, during both symbolic and non-symbolic tasks.

Supplementary material

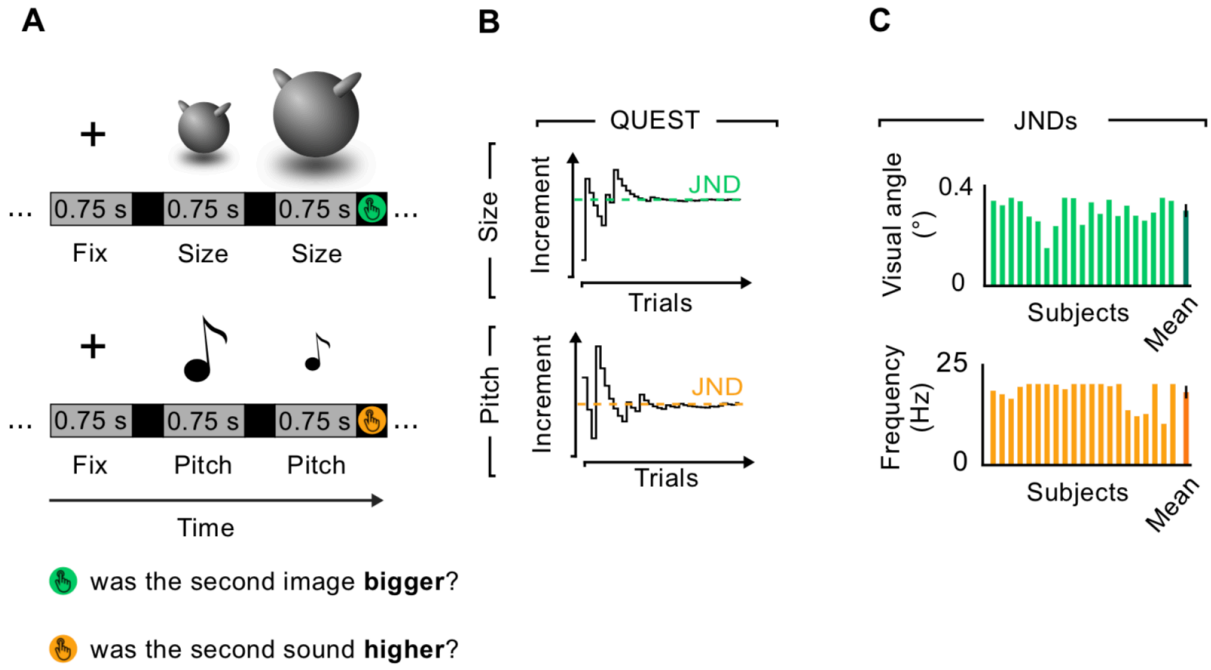

Figure 2.S1 - Psychophysical validation. A-B. Two independent tasks are utilized to extract the Just Noticeable Difference for each participant, using QUEST (Watson and Pelli, 1987). **C.** Participants have very different perceptual sensitivities, and each perceptual space is created on their individual scores.

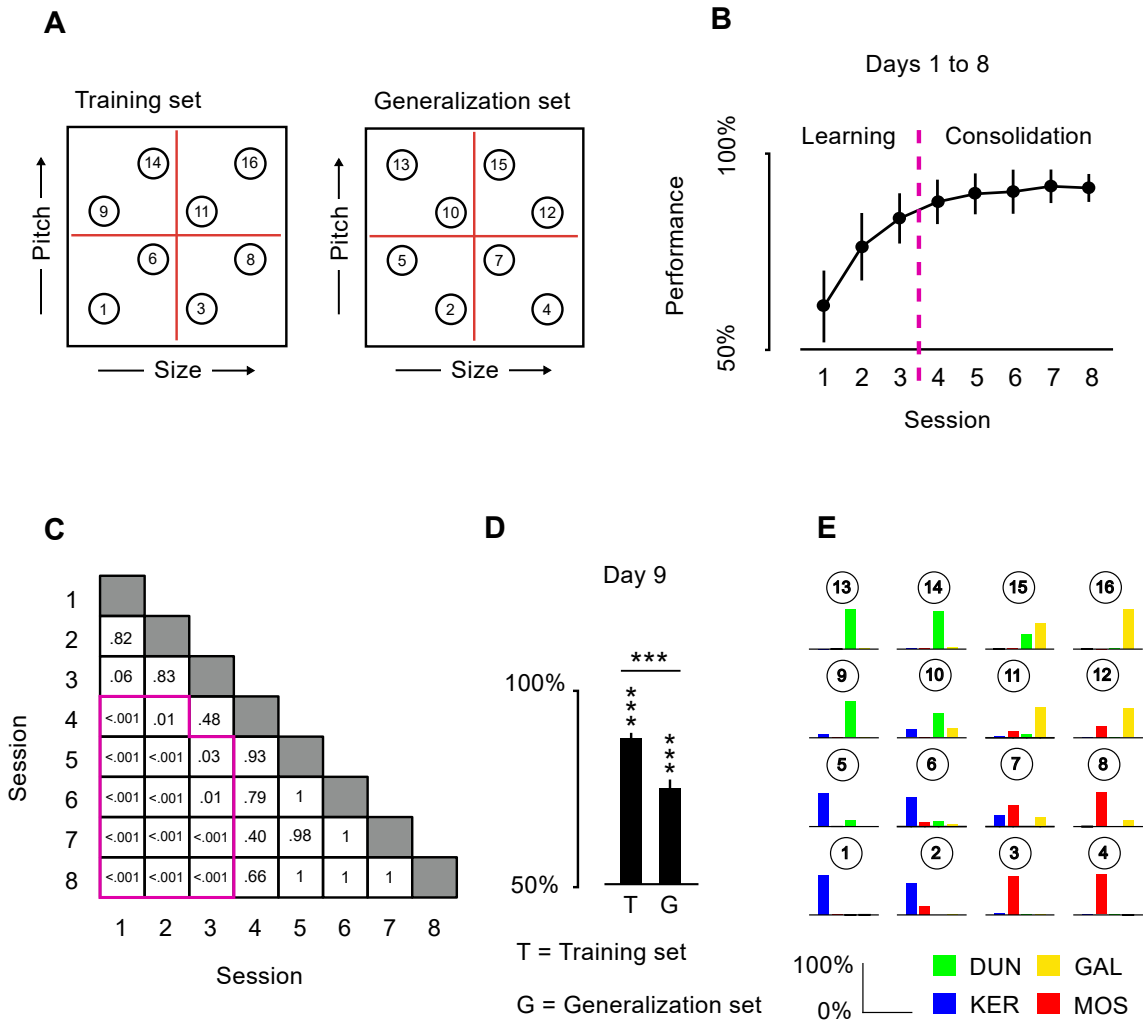

Figure 2.S2 - Training results. A. Different objects are used to train participants and to test for generalization. **B-C.** The learning curve suggests an initial period of learning (sessions 1 to 4) and a subsequent period of consolidation (5 to 8) **D-E.** Performance on the last day (test day) reveals that subjects not only learned to correctly recognise the familiar objects, but also novel combinations, indicating a generalization of the categorical multisensory space that represents the meaning of the novel words.

NAVIGATING A NOVEL SEMANTIC SPACE WITH DISTANCE AND DIRECTIONAL CODES IN THE HUMAN BRAIN

Introduction

Humans and animals have a remarkable ability to orient themselves in space. When we leave our office after a busy working day, we can effortlessly find our way home in the myriad of roads of our city and, in the unfortunate case that the typical route we take is stuck in traffic, we can take a different path to the same destination by surveying the memory we have of the city and selecting an alternative route. The ability to adapt our behaviour in such a flexible way derives from the fact that we stored, in our memory, the knowledge of the city, in terms of where locations of interest are, how distant they are from each other and, by consequence, what are the possible pathways that connect them. In the late 40s Tolman (1949) observed a very similar ability in rats navigating an experimental maze, and he coined the term “cognitive map” to refer to the internal knowledge that they acquired about the experimental setting and the relationships between its elements, such as the distances between different locations or the position of corners, that enabled the animals to easily find shortcuts or alternative routes when obstacles blocked their way to the reward. The neural bases of this spatial knowledge have been described in the following decades, when the hippocampal formation and surrounding areas (for instance retrosplenial cortex and medial prefrontal regions) have been proven to contain spatially-tuned neurons, such as place-cells (O’Keefe & Dostrovksy 1979) and grid-cells (Hafting et al. 2005) that, taken together, contribute to the representation of the animal’s current location and its memory of the surrounding environment. The so-called grid cells, in particular, fire at multiple locations covering the entire navigable surface with a precise triangular periodicity, and are thought to contribute in estimating the distances between points of the physical space to construct a spatial cognitive map (Bush et al. 2015). Crucially, place- and grid-cells have been later observed in humans, during virtual-reality

spatial navigation, using both intracranial recordings (Jacobs et al. 2013) and fMRI (Doeller et al. 2010), not only in the hippocampal formation (mostly entorhinal cortex) but also in medial prefrontal regions, known for their contribution to associative learning and spatial memory and for their strong connections with the hippocampus (see Preston & Eichenbaum 2013 for a review).

A recent proposal posits that humans might use the same neuronal machinery to support an internal representation of non-spatial memories and experiences, recruiting the same neural codes to organise their abstract knowledge in a "cognitive map" of concepts (Behrens et al. 2018; Bellmund et al. 2018). This proposal comes from a set of complementary observations. First of all, the same brain regions where spatially-tuned neurons have been recorded (mostly hippocampus, entorhinal cortex, and medial prefrontal cortex), are also activated during non-spatial tasks, supporting abstract decision making (Schuck et al. 2015; Schuck et al. 2016; Kaplan et al. 2017), and representing, for instance, temporal sequences (Eichenbaum 2014), social spaces (Tavares et al. 2015, Kaplan & Friston 2019), and associative and hierarchical spaces (Dusek and Eichenbaum 1997; Heckers et al 2004; Zeithmanova et al. 2012). Second, in very particular experimental situations, these areas show a 6-fold periodic modulation of their BOLD signal that is consistent with the one observed during spatial navigation - most likely originating from grid cells activity (Doeller et al. 2010) - but during tasks that bear little, if any, similarity with navigation of physical environments, such as imagined navigation (Bellmund et al. 2016), visual search (Julian et al. 2018; Nau et al. 2018), or processing of morphing objects in a 2D visual space (Constantinescu et al. 2016). Finally, in rats and monkeys, lesioning or interfering with mPFC or the hippocampal formation prevents animals to perform tasks where behavioural flexibility is required, such as learning the abstract structure of the task, adapt to reversal learning rules, generalize their knowledge through transitive inference, or finding new shortcuts within a maze (Dusek and Eichenbaum 1997; Buckmaster et al. 2004; Walton et al. 2010; Takahashi et al. 2011; Jones et al. 2012; Koscik and Tranel 2012; Gilboa et al. 2014; Wikenheiser eand Schoenbaum 2016). Taken together, these findings led to the proposal of a shared neuronal machinery for both spatial navigation and high level, concept based, cognition.

But what does it mean to have a “cognitive map” of concepts? A similar intuition had emerged in the fields of neurolinguistic and computational linguistic, where scholars tend to interpret concepts (usually represented by words) as regions or points in an internal space (semantic space), with proximities reflecting similarity in meaning (e.g., see Borghesani and Piazza, 2017). If the idea of conceptual spaces is more than a metaphor, and if the human brain uses the same neuronal machinery used to represent and navigate in physical space to represent and navigate complex conceptual spaces, two predictions follow. First, the activity of those brain regions involved in representing the “cognitive map” of the conceptual space relevant to the task at hand, should also reflect the actual pattern of distances between concepts, when they are considered as regions of an abstract space (the “cognitive space”) which coordinates are the dimensions, or features, that define those concepts. Second, moving between concepts in memories should involve the same direction-dependent neural codes observed in lower level mammals and humans when they navigate the physical space. To date, evidence supporting these two predictions for human conceptual knowledge is still missing, mostly because it is extremely difficult to reduce complex human conceptual representations to low dimensional spaces that allows a comparison with the navigable physical environment.

Here I was interested in studying whether and where the human brain holds a cognitive map of complex concepts, by representing their mutual distances. Additionally, I seek for evidence of directional coding that could be complementary to a distance code. There are at least two definitional criteria for human-like conceptual representations as we intend, and use, in everyday life: that of referring to objects/events classes, or categories, and that of being accessed and manipulated using symbols, such as words or numbers. Because well known concepts are multidimensional in nature and their multiple semantic properties might interfere with our scope, I applied this approach to the same set of data described in Chapter 2, where we created a novel, artificial, but highly controlled conceptual space composed by audiovisual objects that are divided into four categories by means of linguistic labels (words) (Figure 3.1A-B)(see Methods). Twenty-

five adult participants learned to assign each one of the novel objects to a particular category during 9 days of behavioural training. Before and after learning, they were presented with pseudorandom sequences of objects and words and, after learning (because before learning it was impossible) they were asked to bear in mind the conceptual identity of either the object or the word to perform a one back categorization task (see Methods). I reasoned that, for a cognitive map to exist in a brain region, the activity evoked by stimuli referred to different regions of the semantic space should reflect the *distance* existing between them: the closer two concepts are in the semantic space, the closer (or similar) their representations should be. Additionally, I reasoned that subsequent presentations of words and objects referring to different categories implied a specific *direction* travelled within the conceptual space. Therefore, my data were suitable for investigating the existence of both a distance and a directional representational code underlying the representation of a semantic space (Figure 3.1C)(see Methods).

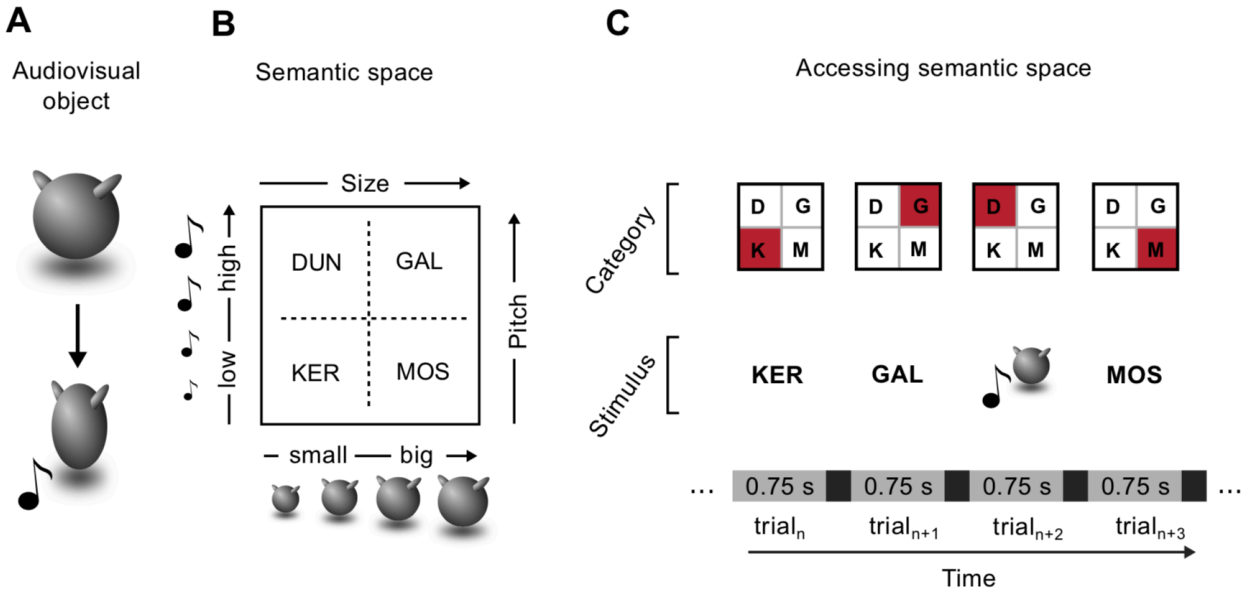

Figure 3.1 - Methods. A. Exemplar of audiovisual object. **B.** 16 multisensory objects are divided into 4 categories by means of abstract words: this creates a novel multisensory semantic space. **C.** Subsequent presentations of either objects or words imply a movement between the regions of the semantic space. These movements cover a certain distance, and have particular directions.

Methods

Participants. You can refer to Chapter 2, section Methods-Participants for identical procedures on the experimental sample.

Conceptual space. You can refer to Chapter 1, section Methods-Conceptual space for identical procedures on how the conceptual space was created. For the current Chapter, the relevant figures are Figure 2.1 (A-B).

Stimuli presentation. You can refer to Chapter 1, section Methods-Stimuli presentation for identical procedures.

Experimental design. You can refer to Chapter 2, section Methods-Experimental design for identical procedures.

fMRI tasks. You can refer to Chapter 2, section Methods-fMRI tasks for identical procedures.

Behavioural Training. You can refer to Chapter 1, section Methods-Behavioural Training for identical procedures.

Neuroimaging acquisition and Preprocessing. You can refer to Chapter 2, section Methods-Neuroimaging acquisition and Preprocessing for identical procedures.

Adaptation analysis. First of all, I assessed what brain regions, after learning, represented the reciprocal distances between the four concepts. I did that by means of adaptation, reasoning that, under the cognitive map hypothesis, a large distance (e.g. from KER to GAL) travelled in the conceptual space should result in an higher release from adaptation compared to a small distance (e.g. from KER to MOS). Functional images for each participant individually were analysed using a

general linear model (GLM). For each run, 14 regressors were included: 1 regressor for each pair of trials of no interest where no movement happened (e.g. two subsequent stimuli referring to the same category); 2 regressors of interest modelling pairs of trials where either a small or a large movement happened in the conceptual space (Figure. 3.2A); 1 regressor indicating that there was a change in the presentation modality (from object to word or viceversa); 1 regressor for motor response; 6 regressor for head-movements (estimated during motion correction in the pre-processing); 3 regressors of no interest (constant, linear, and quadratic). Baseline periods were modelled implicitly, and regressors were convolved with the standard HRF (without derivatives). A high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. I applied group-level analysis within SPM to find brain regions showing a significant adaptation effect for trials where the movement covered a large distance over those where the movement covered a small distance (Family-wise error (FWE) correction for multiple comparisons at cluster level was applied at $\alpha = 0.05$).

Distance-based Representational Similarity Analysis (RSA). In the brain region individuated, at the group level, from the adaptation analysis, I implemented multivariate pattern analysis (MVPa), which is complementary to adaptation. To do that, I run a second GLM. For each run, 22 regressors were included: 1 regressor for each one of the 8 multisensory objects (resulting in 8 regressors); 1 regressor for each one for he 4 words (resulting in 4 regressors); 1 regressor for motor response; 6 regressor for head-movements (estimated during motion correction in the pre-processing); 3 regressors of no interest (constant, linear, and quadratic). Baseline periods were modelled implicitly, and regressors were convolved with the standard HRF without derivatives. A high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. I thus obtained one beta map for each stimulus (object or word) and run. I used these beta maps to conduct a model-based RSA (Kriegeskoorte et al. 2008). I averaged the beta maps for all the stimuli that belonged to the same concept (e.g. two objects that are a KER and the word “KER”) and I extracted, from the Region of Interest (ROI) obtained in the previous analysis, the neural dissimilarity matrix (DSM, 1-Pearson’s correlation) to reveal their distances in the multivariate

representational space. Next I correlated the Fisher transformed DSM to the predicted matrix representing the distances between our concepts (Figure 3.2C). As a control, I repeated the same analysis using pre-learning data, when participants did not have any knowledge of the conceptual space and therefore the distance model should not correlate with the neural data. Finally, to further confirm that no other brain region represented the distance between our novel concepts in the multivariate activity pattern, I run a whole brain searchlight: a sphere was centred in every voxel of the subject- and session-specific datasets, following previous searchlight studies (Connolly et al. 2012). Within each sphere I conducted the same model-based RSA previously conducted in the ROI. I used SPM to test for group level effects, after subtracting the results of two additional searchlights (with matching parameters) that used model-based RSA to look for brain regions responding to differences in either size or pitch, respectively, of the multisensory objects: this was a necessary step to exclude that the multivariate correlation score I obtained was explained by a low-level perceptual coding of differences between objects. Family-wise error (FWE) correction was applied at $\alpha = 0.05$ to correct for multiple comparisons at cluster level. Additionally, I used the neural DSM to reconstruct, using multidimensional scaling as implemented in MATLAB, the most faithful bidimensional representation of the conceptual space, to visualize the spatial arrangement of the four concepts starting from real neural data, and I did it for both pre and post learning datasets.

Direction-based RSA. Next, I asked whether BOLD activity evoked during the transition between two stimuli referring to different concepts was modulated by the direction of the movement in the conceptual space. To do that, I first run a third GLM, now modelling the directions of movement between concepts. For each run, 20 regressors were included: 8 regressors corresponded to the 8 possible directions of movement within the conceptual space, arbitrarily referenced to the horizontal axis; 1 regressor modelling subsequent presentation of two stimuli that belonged to the same conceptual region, corresponding with no movement across the conceptual environment; 1 regressor for changes in presentation modality (e.g. from object to word or vice versa); 1 regressor for participants' response; 6 regressor for head-movements (estimated during motion correction in

the pre-processing); 3 regressors of no interest (constant, linear, and quadratic). Baseline periods were modelled implicitly, and regressors were convolved with the standard HRF without derivatives. A high-pass filter with a cutoff of 128 s was applied to remove low-frequency drifts. I thus obtained one beta map for each movement direction for each run. In the ROI defined by previous analysis, I applied an extension of the similarity-based multivariate approach of Bellmund et al. (2016) to test for the existence of a hexadirectional code in our data, most likely originating from the activity of grid-cells (Doeller et al. 2010). Two movement directions φ and φ' in the interval 0°-359° can be expressed as more or less similar in a n-fold periodic space, by calculating mod(φ-φ', θ), where θ indicates the angle of the periodic (n-fold) grid for which I want to test the modulation. In the case of a grid-like signal, corresponding to a 6-fold periodicity, θ = 60°, and therefore two directions perfectly aligned with a periodicity of 60° would have mod(φ-φ', 60°) = 0. However, if the two directions are not perfectly aligned in the n-fold symmetry, the result of the mod() function indicates the angular distance to perfect alignment. I computed all the 8x8 pairwise comparisons between our sampled movement directions to obtain a model of their predicted 6-fold dissimilarity, corresponding to the angular deviation in the 60° periodic space (Figure 3.3A-B). Next, I applied model-based RSA correlating the 6-fold model to the Fisher transformed neural dissimilarity matrix (DSM) constructed by computing similarity distance (1-Pearson's r) between any pair of distributed activity patterns in the ROI. I computed the correlation between neural data and the model using Pearson's r. To investigate whether this modulation was detectable at the whole brain level, I used the CoSMoMVPa toolbox (Oosterhof et al. 2016) to implement this analysis in a whole-brain searchlight to find cortical regions responding to the 6-fold rotational symmetry. A sphere was centred in every voxel of the subject- and session-specific datasets, following previous searchlight studies (Connolly et al. 2012). Within each sphere I conducted our model-based grid-RSA, storing the resulting correlation score in the center voxel, as summary of the information for the surrounding sphere. To control for potentially competitive periodicities, I applied the same technique within the resulting ROIs, now using periodic models with 4-, 5-, or 7-fold symmetries.

Results

The main goal of the study was to investigate whether and where a (cognitive) map of our novel, multisensory conceptual space would be represented in the brain of adult and healthy participants. I reasoned that the key ingredient of such a map would be to reflecting the patterns of distance existing between the locations (the concepts) of the space it represents. Additionally, I asked whether I could observe a modulation of BOLD signal as a function of the direction of movement in the conceptual space. To test whether any brain region holds these representations, I used a combination of univariate (adaptation) and multivariate (RSA) techniques (see Methods).

Behavioural results. During the learning phase outside the scanner, participants were trained for 8 daily sessions with 8 multisensory objects, performing a delayed match-to-category-name task (see Methods). During the last training session, and without being notified, they were also presented with 8 novel stimuli that they never saw before. These consisted in specific combinations of size and pitch that were absent in the training set, and they were introduced to verify the emergence of a real categorical representation of the semantic space, and not of a mere association between names to individual exemplars. The learning trajectory indicated a significant increment in performance from session 1 to session 8 (session 1: 60 ± 18%; session 8: 89 ± 8%; paired t-test: t24= 8.58, p = 8.86x10-9). Performance collected on session 9 confirmed the successful learning and generalization of the categories (performance training set: 87 ± 7%, different from chance t= 40.29, p = 1.48 x 10^{-23}; performance generalization set: 73 ± 11%, difference from chance, t= 21.49, p = 3.45 x 10^{-17}).

After learning, participants underwent an fMRI session performing a one-back-category-name task (see Methods). Performance in the scanner was high (hit = 84 ± 10%, correct rejection = 97 ± 1%). No participant reported, at the end of the experiment, to have explicitly memorized the stimulus space in any kind of spatial arrangement.

Neuroimaging results

Distance dependent adaptation. First, I investigated whether and where a cognitive map of the novel conceptual space was represented in our participants' brains. I reasoned that a subsequent presentation of two stimuli belonging to two difference categories would cause an adaptation of the BOLD signal that would be proportional to the distance between the two concepts in the two-dimensional concept space. Given our stimulus space we had two levels of distance: a small (e.g. KER preceded by MOS) or a large (e.g. KER preceded by GAL) one. Regions where the BOLD signal is affected by this difference should be detectable using fMRI adaptation, where the response to a given stimulus should be lower when the stimulus is preceded by another one with a small conceptual distance compared to a large distance. By applying this analysis at whole brain level on post-learning fMRI data, I revealed a significant cluster in medial Prefrontal Cortex (mPFC) ($MNI_{x,y,z}$= 3 47 -4; T=6.73; FWE corr.)(Figure 3.2B). No significant cluster was found using pre-learning data.

Distance-dependent RSA To confirm this result with an independent and complementary measure, I extracted from the mPF cluster of the previous analysis the distributed activity patterns for each stimulus after running a second GLM (see Methods). Next, I applied model-based RSA (Kriegeskorte et al. 2008) by correlating the Fisher transformed neural dissimilarity matrix (DSM, 1-Pearson's r) to a model of the predicted distances in the conceptual space (Figure 3.2C). I observed that the multivariate activity evoked in the mPFC post-learning significantly correlated with the model of predicted distances (t=2.78; p = .005, one tail t-test). Again, this was not the case before learning (Figure 3.2D). Additionally, to verify whether this multivariate signal existed also in other brain regions, I implemented our model-based RSA within a whole-brain searchlight (radius of the sphere = 3 voxels, consistent with Connolly et al. 2012) after excluding brain regions responding to differences along either size or pitch between the audiovisual objects (see Methods). I found two significant clusters: one in mPFC ($MNI_{x,y,z}$= 6 50 -10; T=5.71; FWE corr.)(Figure 3.2D) and one in the precentral gyrus ($MNI_{x,y,z}$= 57 5 24; T=6.55; FWE corr.). Given our previous

adaptation analysis and the previous studies indicating mPFC as holding a representation of the cognitive map of task relevant information beyond spatial navigation (e.g. Constantinescu et al. 2016, Shuck et al. 2016), I focused on this region for following controls: first of all, the distance effect in this area was not present before learning (t=-0.2; p = .58, one tail t-test) and, second, the multivariate signal of a spherical ROI constructed around the peak of the searchlight (see Methods) was enough to recover a faithful bidimensional representation of our novel conceptual space (Figure 3.2E).

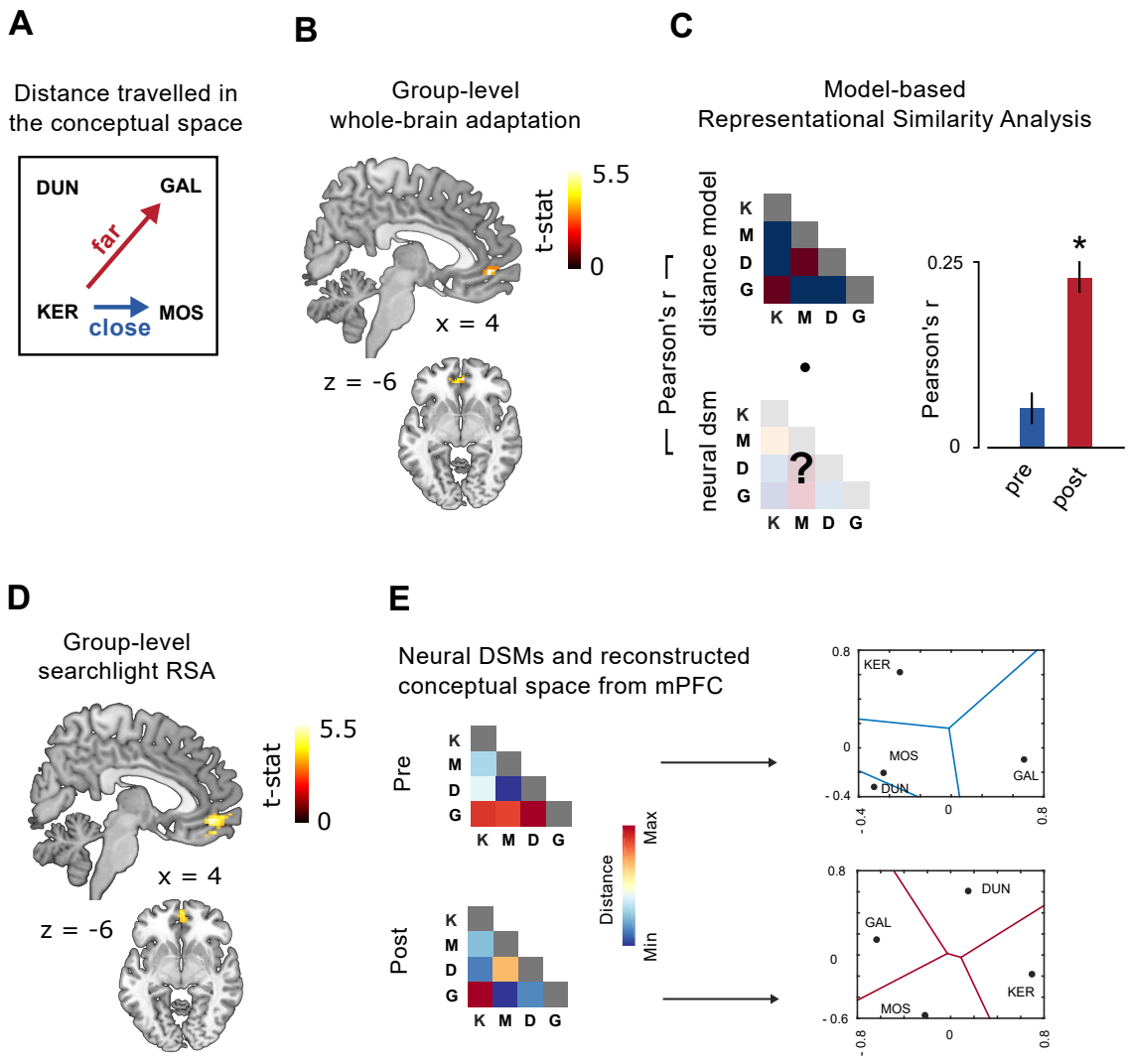

Figure 3.2 - Results of the distance analysis. A. Moving in the semantic space implies covering different distances. **B.** Results of a whole brain adaptation reveal a significant cluster in mPFC reflecting distances between semantic regions. **C-D.** This effect is further confirmed using a independent multivariate approach (RSA) both within the same ROI and with a whole-brain searchlight. **E.** For illustrative purposes, I show the neural DSM both before and after learning in mPFC, which is sufficient to recover a faithful bidimensional representation of the semantic space using MATLAB multidimensional scaling.

Direction-dependent RSA In a previous study (Constantinescu et al. 2016) it was observed that the BOLD signal in the mPFC was modulated following a 6-fold periodicity, typical of grid-cells (Hafting et al. 2005; Doeller et al. 2010), when human participants processed morphing bird shapes varying continuously in their neck:legs ratio, therefore mimicking a movement in an artificial bidimensional "bird space" akin the real-world physical space. Although with our design I could only sample 8 movement directions, I asked whether I could observe a similar modulation between our four discrete concepts, that met both the definitional criteria of human-like, high-level conceptual representations: that of being categorical, and that of having a meaning that is conveyed with a linguistic symbols, such as a word. I fit a new GLM to extract a beta series for each movement direction across the 4 regions of the novel conceptual space (see Methods), resulting in 8 sampled directions. By building on findings that an hexadirectional code can be observed in multivariate activity patterns under highly controlled circumstances (Bellmund et al. 2016), I combined grid-analysis with model-based RSA (Kriegeskorte et al. 2008) (see Methods) (Figure 3.3A). By computing all the dissimilarity measures between directions, I obtained a model (matrix) of the relative distances to the hypothetical hexadirectional grid (see Methods)(Figure 3.3B). Using model-based RSA I could test whether or not the neural dissimilarity matrix extracted from mPFC fit with the model, by computing Pearson's r. This was not the case (t=0.41; p = .68). However, when I implemented the same analysis in a whole-brain searchlight, I did find a set of brain regions where this direction-dependent modulation of multivariate signal was present: right entorhinal cortex ($MNI_{x,y,z}$= 30 5 -32), left orbitofrontal cortex ($MNI_{x,y,z}$= -15 44 -20), left superior frontal gyrus ($MNI_{x,y,z}$= -30 23 60), precentral gyrus ($MNI_{x,y,z}$= 60 5 6; all p<.005 uncorr.). Given previous studies showing directional modulation of BOLD signal as a function of movement direction in entorhinal cortex during both spatial (Doeller et al. 2010) and non-spatial (Constantinescu et al. 2016; Bellmund et al. 2016; Nau et al. 2018; Julian et al. 2018) tasks, and given the high proximity of our entorhinal peak ($MNI_{x,y,z}$= 30 5 -32) to the one reported by Doeller et al. (2010) during spatial navigation ($MNI_{x,y,z}$= 30 3 -30), I focused on this region for subsequent analyses. First of all, I verified that other biologically implausible periodicities (4-fold, 5-fold, and 7-fold) did not account for the signal in this region, and this was not the case (4-fold: t=0.35; p = .94; 5-fold: t=-2.37; p = 0.02;

7-fold: t=-2.14; p = .04). Second, I verified that the 6-fold modulation was not present before learning (t=-0.55; p = .58). Finally, motivated by theoretical and simulation works showing that grid-cells activity can be used to estimate distance between spatial locations to subserve navigation and path integration (Bush et al. 2015), I applied our model-based RSA for distance effect (see above) in the entorhinal cortex. Although I run a whole brain searchlight, I reasoned that, in light of the well-known signal drop in the medial temporal lobe (Schimdt et al. 2005; Bellgowan et al. 2006; Olman et al. 2009), such a distance effect could have been overlooked by our whole-brain correction. Indeed, I did find a weak, although significant, distance effect in this area (t=2.14; p = .02, one tail t-test) that was not present before learning (t=0.48; p = 0.62, one tail t-test)(see Figure 3.3D for the reconstructed bidimensional space using multidimensional scaling).

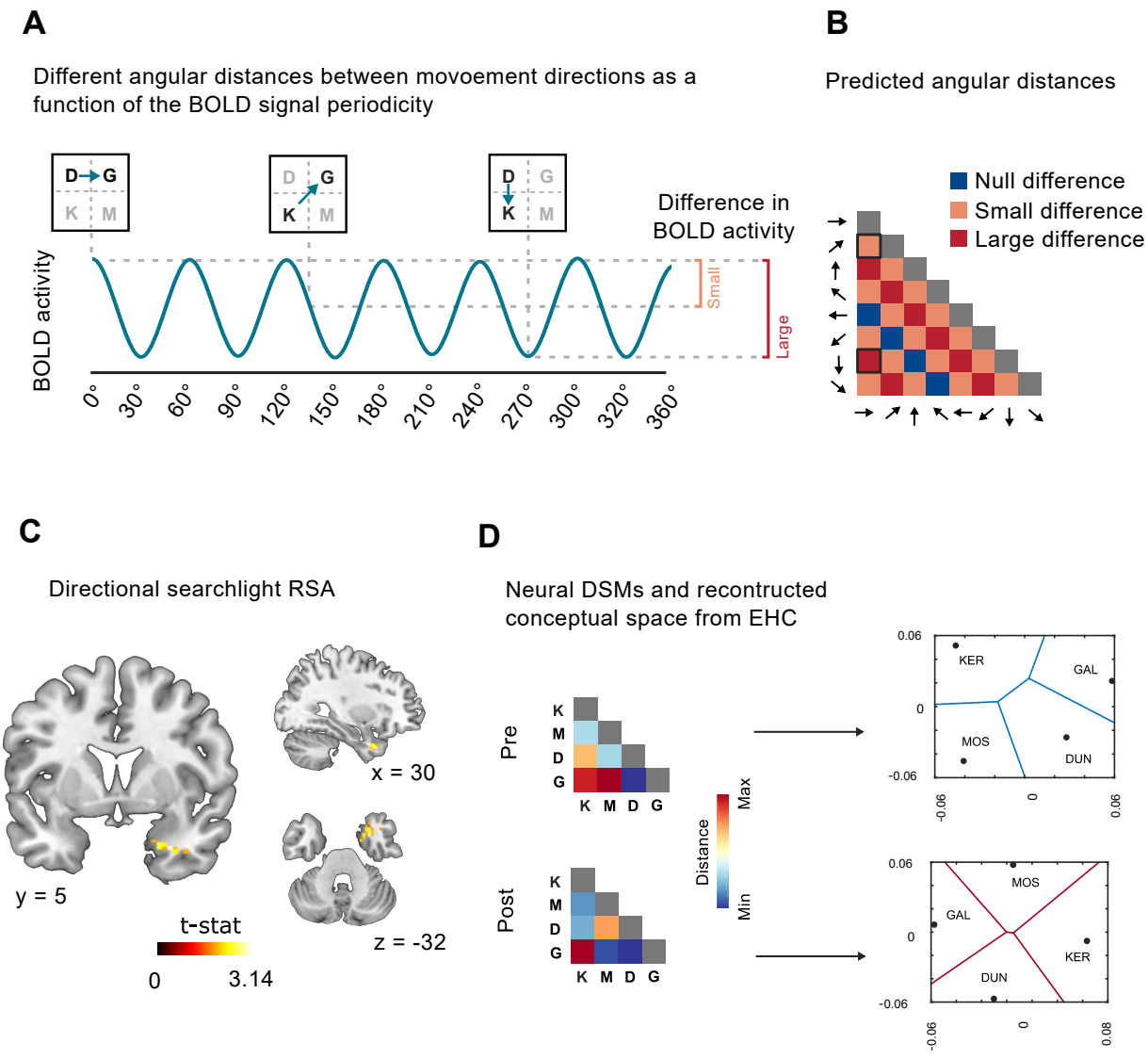

Figure 3.3 - Methods and results of the directional analysis. A-B. Different movement directions elicit multivariate activity that can be expressed as more or less similar in a n-fold periodic space. By applying the function mod(X,Y,n) where X and Y are the two directions and n indicates the periodicity, one can construct a dissimilarity model of the predicted angular distances in the n-fold representational space. **C.** Results of a whole brain searchlight using model-based grid-RSA with the 6-fold periodic model. A significant cluster appears in right EHC. **D.** Distance analysis and illustration of the effect in EHC.

Discussion

In this experiment, I used fMRI to test the two key predictions underlying the "cognitive map" theory: i) that a distance-based representational code, known to represent the reciprocal distances between spatial locations, also represents the patterns of distances between concepts, and ii) that a direction-based representational code, known to be recruited in mammals during spatial navigation tasks, is also recruited when humans navigate among these concepts in memory. Using a training approach allowed me to master the precise metrics of a conceptual space (made of 4 orthogonal categories of labelled multisensory objects). This allowed me to investigate the modulation of the BOLD signal as a function of movement distance and direction between regions of this conceptual space. Through both a whole-brain adaptation analysis and a whole-brain RSA analysis, I observed a strong and reliable distance code in the medial Prefrontal cortex (mPFC). Here, I demonstrated that both the univariate signal and the multivariate activity pattern encoded information about the distance between the four concepts of our novel conceptual space. The distributed pattern of neuronal activity was indeed sufficient recover a faithful bidimensional representation of the conceptual distances. Therefore, my results demonstrate that the activity of the mPFC mirrors the patterns of distance between novel concepts, at least during a categorization task. Additionally, I observed an intriguing modulation, as a function of movement direction, of the activity of the right entorhinal cortex, where grid-like signals have been previously reported in humans using both intracranial recordings (Jacobs et al. 2013) and fMRI (Doeller et al. 2010) during navigation of virtual reality environments. Consistent with simulation studies indicating that grid-cells activity contributes to the estimation of distances between locations in the physical space (e.g. Bush et al. 2015), I found a distance effect also in this region. These results might be

complementary to the distance code observed in mPFC, although some potential limitations of the grid analysis (see below) should be taken into account.

A cognitive map of concepts in mPFC

A previous study by Constantinescu et al. 2016 reported a direction-based modulation of BOLD fMRI signal in vmPFC/OFC (and entorhinal cortex, see below) when participants process previously unknown visual stimuli depicting bird shapes that varied in their neck:legs ratio. This was taken as evidence in favour of an internal "cognitive map" of conceptual knowledge, that allows navigation through concepts as they were locations in the physical space. While the notion of "concept" in this seminal study was taken to indicate a conjunctive representation of two object characteristics, it remained silent on two key features that define conceptual representations in humans: the first one is that in our species concepts typically refer to objects/event classes, or categories, and the second one is that humans construct, recall, and manipulate these categories using linguistic symbols: words. The resulting representations - typically referred to as "semantic representations" - can be conceived as regions in a representational space where the different dimensions represent different features of the objects that the words refer to, the semantic space (Borghesani & Piazza 2017).

In the current work I implemented a training study aimed precisely at extending these previous results and more directly testing whether a distance code underlies one of the most complex of human functions, that of organizing concepts and categories using words. Similarly to Constantinescu et al. (2016), I worked on a highly controlled 2D stimulus set where objects resulted in the orthogonal combination of size and pitch, controlling that variations along these two features were perceptually matched across subjects, using a subject-specific scale (see Methods). However, contrary to them, I engaged participants in a symbolic categorization task, where they had to learn to parse the object space into 4 categories using words, basing on the conjunction of their size and their pitch. By associating more object exemplars to a single category name, participants were constructing true semantic representations, meeting all the definitional criteria of

human-like concepts: that of originating from an arbitrary conjunction of features (similarly to Constantinescu et al. (2016), but here extended to the multisensory domain), that of being categorical (thus defining more broad regions that support generalization, as I demonstrated on the last training day), and that of being labeled with words.

My study also departs from the work of Constantinescu et al. (2016) on the analytical approach used to test for the existence of a cognitive map of concepts. Instead of focusing on directional coding, I first applied a series of univariate and multivariate analyses to find what brain regions would represent the distances between the novel concepts, as the first key ingredient of any map is to reflect the existing distances between locations in space it represents. I did find strong evidence of such signal in the medial region of the prefrontal cortex (mPFC), which has been previously associated to a great variety of cognitive functions different from spatial cognition, such as reversal learning (Jones and Mishkin 1972; Fellows and Farah 2003; Hornak et al. 2004; Izquierdo et al. 2004; Walton et al. 2010), emotion control (Bechara et al. 1999, 2000, 2001; Rampel-Clower et al. 2007), or assignment of economic value (Padoa-Schioppa and Assad 2006, 2008). Stalnaker et al. (2015) discussed all these positions and call them into question by showing that they actually accounted for very specific experimental situations. They propose a more general role of this region, specially in its ventral and orbital portions (vmPFC/OFC), in encoding a representation of the "task space", that is an internal representation of the possible states a participant could be while performing a task. Our results are compatible with this view, as they indicated that during a categorization task, a representation of the "task space" was encode in mPFC: in particular, our task space reflected a categorical and labelled multisensory concept space, therefore extending previous findings that limited the evidence of internal cognitive maps to the domain of reasoning and decision making (e.g. Schuck et al. 2015; 2016).

The entorhinal cortex: coding direction and distances between concepts?

I also investigated the existence of a directional code underlying the navigation in our novel conceptual space. While previous studies relied on quadrature filter procedures (e.g. Doeller et al.

2010; Constantinescu et al. 2016; Nau et al. 2018) to estimate the grid angle on a subset of the data and further test its consistency on independent partitions, I combined the multivariate approach first adopted by Bellmund et al. (2016) together with model-based RSA (Kriegeskorte et al. 2008) constructing a potentially more flexible method. The analysis did not show a significant modulation in the mPFC region that emerged in the previous, distance-based, analysis, but it showed an intriguing result after a whole brain searchlight in the right entorhinal cortex, where grid-cells in rats have been originally recorded (Hafting et al. 2005) and where Doeller et al. (2010) found a grid-like signal when human participants navigated a virtual reality physical environment. Although caution should be adopted in making inferences on neural codes following fMRI analyses, our study indicates that during multisensory semantic categorization - the act of classifying multisensory objects using language, and a hallmark of human high-level cognition - the entorhinal cortex might recruit a directional code similar to the grid-like code typically supporting spatial navigation, to represent movements between regions of the conceptual space. Crucially, in the same area I found a weak, although significant, distance code, which is coherent with the idea that grid-cells support path integration and the representation of distances between locations in the space (e.g. Howard et al. 2014; Bush et al. 2015). It's extremely likely that I did not find any signal of the distance code in entorhinal cortex during our whole brain analyses (adaptation and searchlight RSA) because of the well-known loss of signal in the medial temporal lobe (Schimdt et al. 2005; Bellgowan et al. 2006; Olman et al. 2009). It is interesting to notice that I did not observe the opposite pattern in the mPFC: although I found a very strong modulation of its signal as a function of distance, no evidence of directional coding was found in this region. There are at least two possible explanations, which are not mutually exclusive. First of all, although the argument introduced above posits that grid-cells might support the representation of distances between location in a physical, or conceptual, space, this might not be the only way our brain represent distances. Indeed, more recently evolved brain regions such as the mPFC might have developed different representational codes to support the same function. Another possibility is that the two regions - mPFC and entorhinal cortex - both support the representation of a cognitive map of concepts useful to solve the task at hand, but by playing different roles: a possible scenario, in light

of the well-known connectivity patterns between the hippocampal formation and the mPFC (see Preston & Eichenbaum 2013) is that the former, and in particular the entorhinal cortex, informs the latter by using a grid code. In this picture, the distance code reflected in the mPFC would be the result of the computations happening at a lower level in the entorhinal cortex. Future studies should try to address this issue with more specific experimental designs and measures.

Possible limitations

There is actually a third, and potentially more parsimonious reason why I did not observe a directional modulation of the mPFC signal, that could also explain the weaker statistical significance of the directional signal in EHC, which refers to a limitation of our experimental design: a sub-sampling of the possible directions of movements within our 2D conceptual space. In fact, given that I only had 4 categories I could only sample 8 movement directions, that might not offer enough information to properly estimate the grid-signal in areas different from the entorhinal cortex, at least compared to other experiments (e.g. Constantinescu et al. 2016) where, thanks to the availability of a continuous space (which in turn would make less plausible a generalization to discrete human-like conceptual representations), many different directions could be tested. Moreover, due to the same sub-sampling, I should notify that our 6-fold symmetry model is also compatible with the presence of a 2-fold symmetry. While a 6-fold symmetry readily derives from the known hexagonal arrangement of the tuning functions of neurons recorded in the entorhinal cortex and in an extended memory network in humans (e.g., Hafting et al. 2005; Doeller et al. 2010; Jacobs et al. 2013; Bellmund et al. 2016; Constantinescu et al. 2016; Nau et al. 2018; Julian et al. 2018), a 2-fold rotational symmetry would correspond to a population of neurons tuned only to a given direction φ and to its opposite φ + 180°. This response has never been reported to date neither in neuroimaging or electrophysiological works, and thus seems to me as largely implausible. However, while I cannot firmly exclude the presence, in the human brain of a neuronal populations characterized by a two-fold symmetrical tuning function, I can certainly claim that its presence is, based on our current knowledge, biologically rather implausible. However, even in the

case that our results actually reflect a 2-fold symmetry, they would still represent evidence for a directional coding that is characterized by either a 2 or a 6 fold periodicity, but not by other periodicities (4-, 5-, 7-fold), and that underlies navigating abstract conceptual spaces in the entorhinal cortex. Thus, at the very least our results represent novel evidence for the presence of both a distance-based and a directional coding (compatible with either a 6-fold or a 2-fold symmetry) of concepts and movements among them.

Conclusions

To conclude, in humans symbol-dependent categorical format of representations defines behaviourally relevant regions of the knowledge space that we typically refer to as “concepts”. As human cognition critically depends on language, it is essential to encode the relationships between its units (the meaning of the words) to support generalization, abstractions, and inferences, the key elements of human flexible behaviour (Behrens et al. 2018). Our results indicate that the medial PFC encodes these relationships through distance dependent code, and reveal weaker but potentially informative direction and distance dependent modulation of entorhinal signal. On a more general perspective these results may be seen as a novel example of “cortical recycling” (Dehaene & Cohen 2005): brain regions holding specific coding schemes that evolved, in lower-level animals, to represent spatial relationships between objects and locations crucial for spatial navigation, in humans are reused - or “recycled” - to encode relationships between words and concepts in an internal cognitive map (Tolman 1947; O’Keefe & Nadel 1978).

OBJECT NAMING SUPPORTS THE EMERGENCE OF GENERALIZED SEMANTIC CATEGORICAL SPACES

Introduction

Humans are able to acquire, store, and recall much information about what they experience in everyday life. This knowledge we have about things in the world is stored in our conceptual memory, which is organised into behaviourally relevant categories. Categories are equivalence classes, based on highly similar patterns of activation for all the items they are composed by, and range from perception- based equivalences (a dog and a cat shares the typical shape of all the mammals) to more sophisticated behaviourally relevant and abstract similarities (both a pen and a computer can be seen as writing devices). Humans construct categories by means of symbols, and this gives rise to internal meaningful semantic representations.

But what are the advantages, if any, of having conceptual representations of categories that are defined by means of symbols (semantic representations)? Previous studies showed that when participants were asked to find items of a given category (e.g. a number), they were more accurate and faster in giving the correct responses when a linguistic and redundant label, matching with the category of the expected response, was offered as a cue (Lupyan & Spivey 2010; Lupyan & Thompson-Schill 2012). Pierce & Lupyan (2015) argued that this might be due to the fact that symbols and labels refer to abstract representations that do not covary with specific instances of the class or category they refer to, but rather they act as “unmotivated” cues, holding a more general, symbolic or abstract representation of the category. On the contrary, other cues such as the picture of a dog, or the sound of the animal barking, would necessarily convey information about a specific exemplar, as they would refer to the category via an iconic relationship.

If symbols affect the access to conceptual knowledge by enhancing its abstract meaning rather then the representations of specific and individual exemplars, it is reasonable to assume that the same facilitation emerges during the process of acquisition of this knowledge, that is during categorization. Indeed Lupyan et al. (2007) showed that human adults were more proficient in learning to parse novel visual objects, representing fictitious aliens, into categories, when labels representing their category names were provided with them compared to a situation in which no label, or a non verbal label, was given. These effects have been to date long established in both adults and children (Balaban & Waxman 1997; Fulkerson & Waxman 2007; Ferry et al. 2010), and Althaus & Plunkett (2016), following Waxman & Markow (1995) showed that in 12-months-old children the facilitation effect provided by linguistic labels triggers an attentional focus on the similarities between different objects that share the same name.

Here I report the first results of an ongoing study focused on adult subjects to investigate the hypothesis that, besides facilitating categorizing training stimuli, symbols affect the way novel object exemplars, for which participants were not trained on, are categorized: in brief, I asked whether categorical judgements are more generalisable when the categories are constructed using symbols.

I trained 40 adult participants for two days in parsing a multisensory object space into categories. One group acquired categories by means of a non-symbolic, where they had to indicate whether two objects belonged to the same category or not, while the other by mean of a category naming task. Next, I tested both groups on a third day with a non-symbolic task - the same used to train the non-symbolic group. Crucially, half of the trials during the test phase consisted of novel objects, representing previously unseen combinations of features, but falling within specific portions of the the learnt categorical space. By comparing performance during the test day between the two groups and between novel vs old category exemplars, I seek to verify whether symbolic categorization affects later categorical generalization.

Methods

Participants. 40 adult volunteers were recruited for the experiment (eleven males; mean age = 23.6, std = 2.1). All participants gave written informed consent, and were reimbursed for their time. They were right-handed with normal or corrected-to-normal vision.

Conceptual space. You can refer to Chapter 1.3, section Methods-Conceptual space for identical procedures. The relevant figures for the current chapter are Figure 4.1A-B-C.

Stimuli presentation. You can refer to Chapter 1.3, section Methods-Stimuli presentation for identical procedures.

Experimental design. The experiment consisted of two parts: training and test. At the beginning of the training phase, each participant was randomly assigned to either the symbolic (S) or the non-symbolic group (NS). For both groups, the training phase lasted 2 days, but the tasks they were involved in changed on the basis of the group. On the third day of the experiment, participants performed a test task that was identical for both the groups. Figure 4.1D

Training phase. The S-group performed a delayed match to category name task for 2 training days, where they had to learn the correct category name for each multisensory object.
Each training session was approximately 10 minutes long, and it was divided into 4 mini-blocks of 20 trials each, for a total of 80 trials. It started with a brief presentation of the objects as exemplars of the four categories (KER, MOS, DUN, GAL). After this familiarization phase, each trial consisted of an object presentation (750 ms), followed by a fixation cross (500 ms), and by the presentation of one category name. Each object was presented 10 times per training session. Participants were instructed to press one key on the keyboard to select the whether the presented name was the one indicating the correct category of the object. They were asked to respond as fast as possible, but no time limits were imposed. After their response, an immediate feedback appeared on the screen

for 1000 ms, indicating with the words “Correct!” or “Wrong!” the accuracy of the choice. After each miniblock, participants would be provided with the cumulative percentage accuracy. For half of the trials, objects were followed by the correct name, while for the other half they were followed by a wrong name: this led to a chance level of 50%.

The NS-group performed a same or different category task for 2 training days, where they had to learn whether two objects belonged or not to the same category, without the help of linguistic labels.

Each training session was approximately 10 minutes long, and it was divided into 4 mini-blocks of 20 trials each, for a total of 80 trials. It started with a brief presentation of the objects as exemplars of the four categories, without specifying their names. After this familiarization phase, each trial consisted of an object presentation (750 ms), followed by a fixation cross (500 ms), and by the presentation of another object (750 ms). Each object was presented 10 times per training session as first stimulus. Participants were instructed to press one key on the keyboard to select the whether the presented objects belonged to the same category. They were asked to respond as fast as possible, but no time limits were imposed. After their response, an immediate feedback appeared on the screen for 1000 ms, indicating with the words “Correct!” or “Wrong!” the accuracy of the choice. After each miniblock, participants would be provided with the cumulative percentage accuracy. For half of the trials, the two objects belonged to the same category, while for the other half they did not: this led to a chance level of 50%.

For both the training protocols I used a subset of 8 objects (exemplars number 1-4-6-7-10-11-13-16). The remaining 8 objects (exemplars number 2-3-5-8-9-12-14-15) were used on the test phase (see below).

Test phase. During the test phase both groups performed the same non-symbolic same-or-different-category task that the NS-group was trained on. The procedure was exactly identical to the one described above, except for the fact that i) no feedback was given after the response, and ii) now all the 16 objects were used. In particular, each trial could be of two types: type 1 (old-old trials) presented two objects that were both known to participants, because they were selected

from the subset they have been previously trained on; type 2 (old-new) trials presented them with a known object, followed by a novel object they never saw before. Novel objects were constructed with the complementary audiovisual combinations missing from the training object set. This led to a longer session of about 20 minutes. Crucially, participants were not notified about the presence of novel objects.

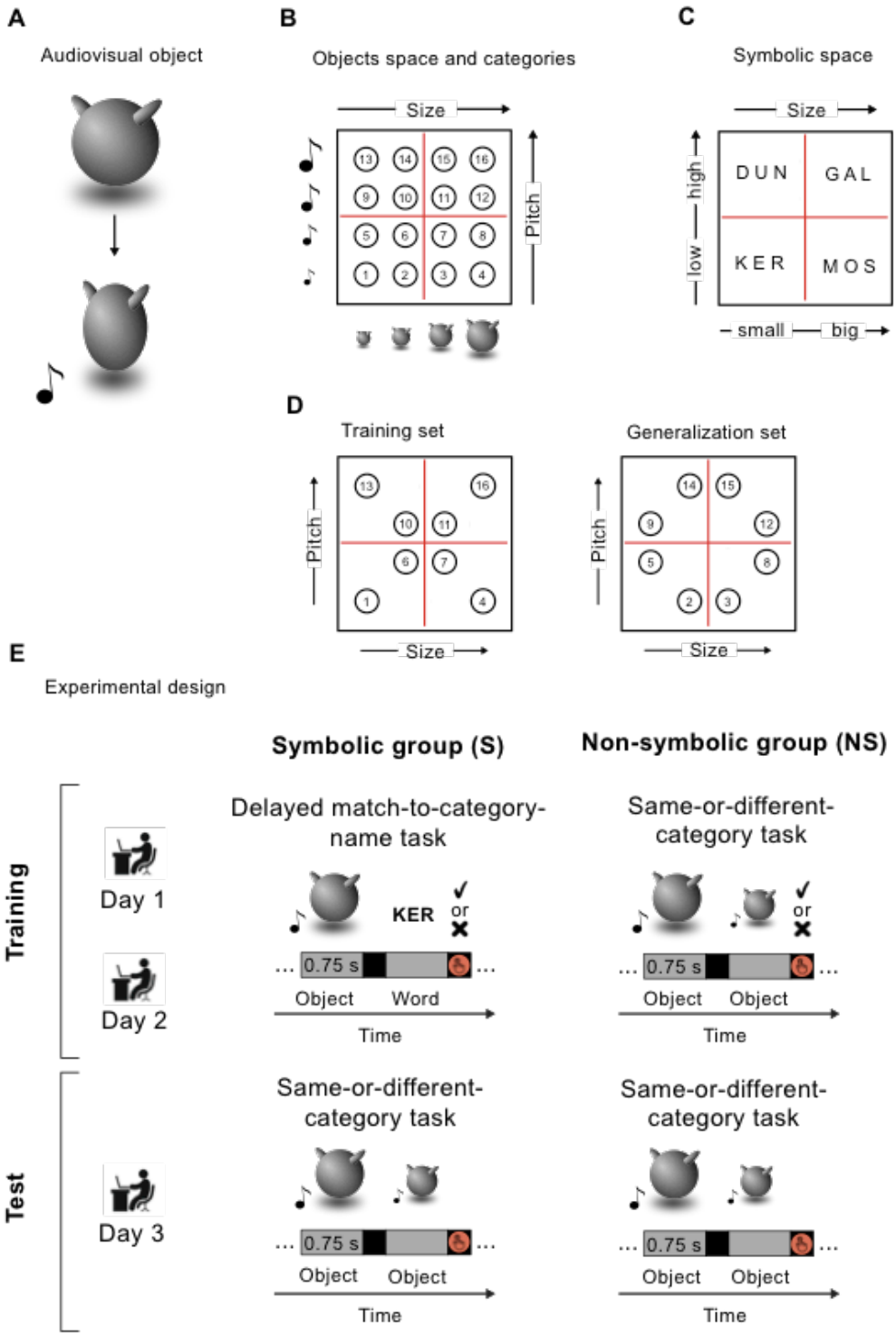

Figure 4.1 - Methods. A-B-C. 16 audiovisual objects are divided into 4 categories by means of symbols. Symbols are revealed only to half of the participants. **D.** a subset of 8 objects is used during the training, while the remaining 8 objects (generalization set) are used, together with the old ones, on the test day. **E.** Half of the participants are trained with a symbolic task (delayed match to category name), while the other half with a non symbolic task (same or different category task). On the last experiment day (test day) they are both tested with a non symbolic same or different category task, but novel objects are introduced together with the old ones.

Results

Learning performance during training

Both the S-group and the NS-group had significantly above chance performance during day 1 and day 2. The S-group significantly improved its performance in object naming from training day 1 to training day 2 compared to the NS-group, as revealed by a significant interaction after a 2x2 repeated measure ANOVA ($F=14.06$, $p = 0.001$) and post-hoc t-tests (see Figure 4.2A).

Generalization during test phase

Crucially, during day 3 (test day) the S-group was able to correctly categorize objects during old-old and old-new trials, while the NS-group was significantly worse with old-new trials, where novel exemplars were introduced (2x2 repeated measure ANOVA, $F=6.006$, $p = 0.024$). Post-hoc t-tests reveal that the effect was driven by the fact that NS-group was worse with old-new trials compared to old-old trials ($t= 2.94$, $p = .008$) and that it was worse with old-new trials compared to the S-group ($t = 2.90$, $p = .009$)(Figure 4.2B).

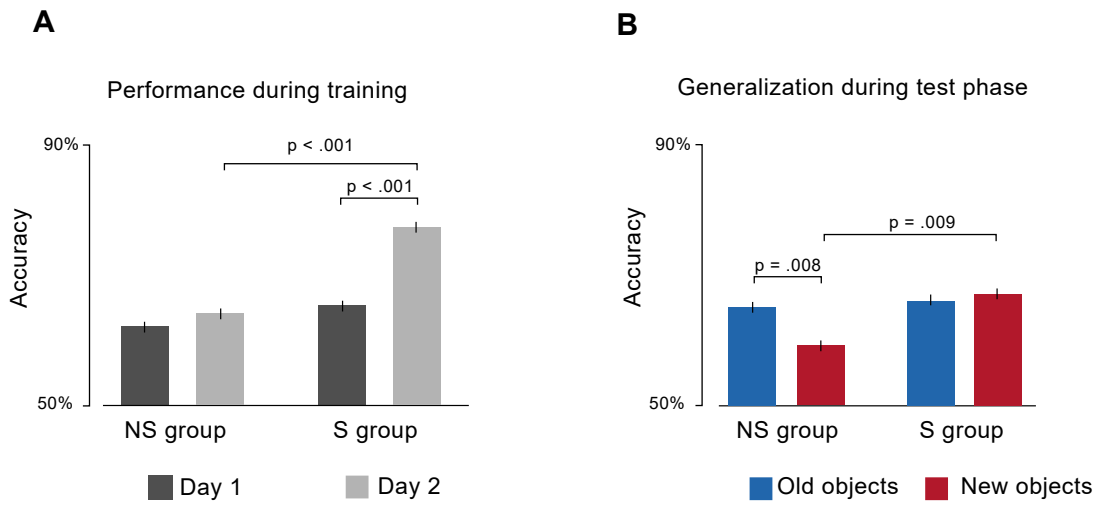

Figure 4.2 - Results. A. The symbolic group (S) significantly improves its performance during the training, from day 1 to day 2. **B.** Crucially, the symbolic group during the test day was able to correctly categorize both trials with only old objects and trials where novel object combinations were presented. The non symbolic (NS) group on the contrary wa significantly worse during trials with novel objects.

Discussion

Here, I have reported the preliminary results of a study aimed at investigating the effects of symbolic categorization on generalization. So far, I have observed that participants who learned to categorize objects by means of linguistic symbols (written words) not only improve their performance during training faster than a control group that learned the categories without symbols, but they extended categorical judgments in a subsequent non-symbolic categorization task also to object exemplars that were new and for that they had not received explicit training. This was not the case for the non-symbolic group, who performed, during the final test, significantly worse when categorical judgments involved novel objects rather than familiar ones.

These results, although limited as the study is still going on at the moment when I write, are of potentially interest for the present dissertation and worth mentioning, because they immediately speak on the nature and function of semantic representations that emerge after categorical learning.

In a previous study, Lupyan et al. 2007 trained adult participants in grouping novel objects (fictitious aliens) as members of two behaviourally relevant categories, basing on their visual appearance, as exemplars to either approach (first category) or to avoid because dangerous (second category). Two groups of people underwent this training performing almost the exact same training procedure for which, after each response, a feedback was given. However, a crucial manipulation was introduced to make the two protocols different: one of the two groups received, together with the feedback, a redundant information on the name of the category that the object just presented belongs to. This simple, presumably irrelevant manipulation, had a singular effect on categorical learning: the symbolic group became more faster and more accurate in correctly categorize objects. These effects were not present in the non-symbolic group, neither in a control group that received a non linguistic cue instead of the word label to identify the categories. These results were suggestive of two important aspects of creating conceptual (categorical) representations using linguistic symbols: first of all, words significantly facilitate the acquisition of the conceptual knowledge they represent, and this finding has been replicated many times, also in children within the first 10 or 12 months of life (Althaus & Westemann 2015; Althaus & Plunkett 2016); second, they indicated that there is something special in words and symbolic labels, that other cues used with the same purpose do not have.

With the present study I wanted to investigate what are the key aspects that make symbols so useful in constructing conceptual representations, and in general what are the advantages of having semantic representations at all. Although preliminary, our results seem to indicate that one crucial aspect facilitated by the presence of words is generalization: categorical representations constructed using symbols are more generalizable than those constructed without them.

Ongoing and future research will have to address this issue more carefully, also try to link these findings to changes in brain activity (see Chapter 2).

GENERAL DISCUSSION

Acquiring knowledge about the world and organizing it into meaningful categories is a complex process, one that the human brain solves with a striking ease thanks to the use of language and symbols, that define the transition from perceptual to semantic representations, and therefore from perceptual to conceptual spaces. The majority of cognitive neuroscience studies on semantic memory focus on well-known concepts. Here I had the unique opportunity to observe how the plastic nature of the human brain supports one of the key mechanisms of knowledge acquisition: learning to recognise individual multisensory objects and to categorize them through the association with arbitrary symbols, thus to create novel semantic representations. Learning-induced changes happened in a set of brain regions, only partially overlapping with the classical "semantic network" (Binder et al. 2009), where both sensory and associative areas modified their representational geometries to support the emergence of stable representations of individual objects and the association of such representations to those of their names. These observations revealed that novel semantic representations, where symbols (words) acquired meaning through the association to a referent (classes of multisensory objects), originated from the orchestrated and integrated activity of distributed perceptual and memory systems, of which the main players are:

- the Visual Word Form Area (VWFA), an area in the left fusiform gyrus best known for its role in word recognition (Dehaene & Cohen 2011), which multivariate activity after learning encodes both words and objects in a way that reflect their categorical association
- the left Angular Gyrus (AG), an area of the left parietal lobule best known for its role in multisensory integration and for being one of the key nodes of the semantic network (see Binder et al. 2011), where the representations of object identities as unique multisensory combinations emerged after learning and where objects and the corresponding categorical names became similarly represented, somehow mirroring the pattern observed in the VWFA;

- the right Hippocampus, a subcortical area in the Medial Temporal Lobe playing a crucial role in associative learning (Suzuki 2007), here representing words and their corresponding categorical names in a way that fully abstracts from the presentation modality, and which activity strongly correlated with behavioural performance in categorization outside the scanner;
- the auditory and visual associative cortices in the Superior Temporal Gyrus (STG) and in Lateral Occipital Complex (LOC), respectively, that after training developed high sensory specificity, supporting sensory segregation.

In Chapter 3 I also showed how these semantic representations, being conceivable as regions of a novel semantic space, are encoded in the medial Prefrontal Cortex (mPFC) and, to a weaker extent, in the right Entorhinal Cortex (EHC) using the same neural codes typically employed during spatial navigation (e.g. Doeller et al. 2010). In particular, I demonstrated, using two independent techniques (fMRI adaptation and Multivoxel pattern analysis), that the mPFC encoded the distances between concepts in the novel semantic space, while the right Entorhinal Cortex, besides encoding distance, also showed traces of encoding the direction of movements in the semantic space, as mimicked by the sequential processing of stimuli. Finally, in Chapter 4 I presented the preliminary results of a behavioural investigation aimed at describing the advantages of using symbols during categorical learning, with a particular focus on generalization: here I demonstrated that categorical judgments can generalize to novel exemplars when categories are learned using symbols, compared to when symbols are not used.

Taken together, these results are, in my view, complementary. The role of the hippocampal formation in semantic knowledge has been traditionally overlooked in favour of its well-documented involvement in episodic memory. The results of i) the cross-modal searchlight (Chapter 2) and ii) the directional-RSA searchlight (Chapter 3) point to this macro-structure in the Medial Temporal Lobe as participating in the process of learning and representing concepts and meanings, although the two analyses revealed different sub-portions of this area: the hippocampus and entorhinal cortex, respectively. It is interesting to notice, in this respect, that post hoc analyses

(that I only report here) indicate that the hippocampus does not show any evidence of directional code (RSA grid analysis t = -1.71, p = .10), confirming previous intracranial recordings in rats (e.g. Hafting et al. 2005), and that the entorhinal cortex did not showed cross-modal similarity between objects and their corresponding names (t = .97, p = .33). This strongly support the existence of independent processes going on in these two sub-regions of the hippocampal formation. Among them, the right Hippocampus was the one that puzzled me the most, mostly because of its lateralization: even lowering the threshold of the statistical tests I did not observed significant results in the left hemisphere for the crossmodal searchlight. The right hippocampus is best known for its role in spatial memory, which should not have any relevance in my task. However, as demonstrated in Chapter 3, brain regions holding specific coding schemes for supporting spatial navigation also recruit them to represent a novel semantic space. Among these, and besides the EHC, I found a strongly significant cluster in mPFC, a region that is highly recognized for its pivotal role both in spatial and non-spatial tasks and for its strong connections to the hippocampus (Preston & Eichenbaum 2013). This raised the fascinating hypothesis that the right hippocampus supports the creation of novel semantic representations by recruiting a spatial code. To test this, I verified whether the the similarity of the neural representations of the four concepts in participants' hippocampus reflected their mutual distances in the underlying semantic space. To do that, I applied the same multivariate approach introduced in Chapter 3, which revealed such a distance-dependent representation in mPFC. In this post-hoc analysis centered in the hippocampus, I found a weak, but significant, result (t = 2.09, p = .023, one-tail). This did not happen in any of the other regions emerging from the crossmodal searchlight showed in Chapter 2. This additional analysis is informative for at least 3 reasons:

i) it suggests more strongly that the right hippocampus might use a spatial code to contribute in associating words to their meanings, potentially by representing the objects that become the referents for the symbols in a spatial format. This would enable answering the question about the lateralized effect reported in Chapter 2 (indeed, a very recent work by Theves et al. 2019 found that the right, and not the left hippocampus encodes distances between the representations of objects arbitrarily assigned to positions of a bidimensional visual space);

ii) the results suggests that a similar computational strategy is employed by the mPFC (although it did not show evidence for cross-modal similarity ($t = 0.20$, $p = 0.84$)), pretty much as it has been shown to do for “task-spaces” in decision making experiments (see Wilson et al. 2014; Schuck et al. 2016) ;

iii) that the idea of a “cognitive map” reflecting conceptual or semantic spaces for the ongoing tasks via the recruitment of spatial-dependent codes (Bellmund et al. 2018) is potentially true, but at the same time it needs a better definition to be able to explain the exact computations underlying the role of these different mid-temporal and frontal areas (and potentially parietal, see Doeller et al. 2010 and Constantinescu et al. 2016).

An intriguing set of results from Chapters 2 and 4 suggests that one of the directions future studies should focus on is the study of the interplay between semantic representations, spatial codes, and generalization: in Chapter 2 I showed a behavioural effect of generalization in naming novel objects after training participants on a different subset of exemplars, and this effect correlated with the crossmodal similarity observed in the right hippocampus. In Chapter 4 I showed preliminary behavioural results that a similar effect happens for non linguistic categorical judgments after symbolic categorical learning. Is there a relationships between the putative recruitment of a spatial code and the advantages in generalization following symbolic categorization? In light of this question, which will be objective of future research, a set of results from intracranial recording studies conducted by the team of Rodrigo Q. Quiroga, in collaboration with Itzkav Fried become relevant. In a sequence of fascinating experiments they reported that individual neurons in the hippocampal formation of human adults represent what they called “concepts” (that are, individual identities of famous people, places, animals, etc.) in a very selective and highly invariant fashion: a neuron responding to the identity of Luke Skywalker would fire only for pictures of that character and not to the ones of other ones but, crucially, it would fire also to other stimuli representing its very same identity, such as his written or spoken name (Quiroga et al. 2005; 2012). It has been claimed that such representations support declarative memory functions specifically in light of their invariance to basic metric details of the stimuli, holding a truly abstract or conceptual representation of the stimulus processed. If this is the case, than these neurons would serve as the

perfect interface for generalization, because they would hold a “core”, abstract representation of a concept that does not reduce to the specific physical characteristics of the stimulus just presented, and such core representation could be used to evaluate whether a novel stimulus is similar or not or, to put it in another way, how much closer it is in the corresponding conceptual space. Recent theoretical works elaborated on the interesting similarities between humans’ concepts cells and place cells observed in lower level mammals (Horner & Doeller, 2017; Behrens et al. 2018), suggesting that both these neural behaviours are signatures of the same underlying (spatial) code.

While the specific other results obtained in my works are discussed in the relative “Discussions” sections for each Chapter, there are some questions that remain unaddressed, and will be object of future attention and research:

1. Despite strong evidence from clinical and neuroimaging studies (see Lambon-Ralph et al. 2017 for a review) I did not find any representation of the novel concepts in the Anterior Temporal Lobe (ATL). This might be due to the stimuli I used, that varied in size and sound, but not in shape. The ventral visual pathway is known to encode shapes of visual objects and stimuli, and some neuroimaging works has tried to disentangle whether object representations in the ventral visual pathway are defined by the visual properties of the pictorial material used (e.g Proklova et al. 2016, but see Proklova et al. 2019). Indeed, a simple yet fundamental conceptual distinction such as living vs. non-living concepts can be fully accounted for by similarities within the “living” category (e.g. most of the animals have similar body structures, with a head, a body, and a number of legs between two and eight) and across the “living” vs. “non-living” categories. This is one of the reasons that motivated ,the few neuroimaging studies that make use symbolic stimuli such, as words, for referring to items in the respective categories. A recent study by Borghesani et al. (2016) for instance demonstrated that the ATL contains conceptual information sufficient for discriminating among different clusters of animals, but it was based on a ROI-based approach that ignored regions outside the ventral stream, in particular the AG. In our study

we found that the left AG represents object categories in a way that is very similar to their corresponding names. The AG lies at the perfect intersection between auditory and visual pathways, making it the ideal candidate as a convergent zone to merge together audio-visual information into more abstract representations. Future studies should address directly the specific contributions of the left AG, the ATL, the VWFA, and the hippocampal formation in representing conceptual knowledge, controlling for the type of task used and for the sensory modalities upon which the investigated concepts are best defined by.

2. Classical accounts of memory consolidation indicate that the hippocampus provides a fast-learning interface for rapid encoding of new memories that are slowly written in the neocortex for later recalling (e.g. McClelland et al. 1995). Our results support this view, and stress i) on one hand the crucial role of the hippocampus in learning and supporting conceptual representations, and ii) on the other hand the role of neocortical regions, such as the VWFA and L-AG, in holding these representations beyond learning. It worth mentioning, however, a recent paper by Brodt et al. (2018) where the authors, by showing learning-induced changes rapidly emerging in human posterior parietal cortex (Precuneus, known to be part of the semantic system) and lasting for more than 12 hours, challenged traditional models of memory consolidation, reporting evidence of a fast emergence of memory traces outside the hippocampal formation. Further studies will have to address the temporal dynamics of conceptual learning in the human brain, monitoring the emergence of semantic representations during multiple learning stages, and how higher associative regions interact with lower sensory cortices during this learning process.

3. The training procedure I designed involved the use of words to tile the perceptual space into categories. I chose this approach because humans construct and organize their conceptual knowledge of the world using language, giving rise to semantic representations where meanings are conveyed by symbols. Lupyan (2007, 2012, 2015) demonstrated that language enables subjects to focus on those sensory characteristics that define objects

and object categories, eventually fostering the creation of concepts. This has been investigated in carefully controlled uni-sensory domains. Human experience, however, is much more articulated and multisensory in nature. What is the role of symbolic categorization in multisensory perception? How does it give rise to complex categorical spaces, or conceptual/semantic spaces? Future studies should address these questions and look for careful descriptions of the effects of symbolic categorization for instance in multisensory perceptual judgments, generalization to novel exemplars, and creation of categories of abstract concepts of different types than objects, such as episodes or events. The investigation described in Chapter 4 is currently going toward this direction.

4. I showed that the mPFC, the entorhinal cortex, and also potentially the hippocampus, mostly known for encoding spatial locations using a variety of spatial codes, also encodes the geometry novel bi-dimensional semantic space. However, human experience of the world is not confined to two dimensions, nor it is possible to reduce all the complexity of our semantic knowledge to bi-dimensional feature spaces (or cognitive spaces) relevant for the ongoing task. To date, a single study attacked the question of whether the hippocampal formation used the same codes for representing non bi-dimensional environment, but it focused on rats, and on 1-dimensional sound spaces (Aronov et al. 2017). Despite interesting, as this task was not spatial in nature, much more work is required to investigate what these cells are actually involved into. Semantic spaces offer the unique opportunity to test not only neural representations that are relevant for humans, but also to test multiple dimensions and their representations beyond 2D. If place- and grid-cells are recruited to encode both 1- and 2-D non-spatial knowledge, as indicated by Aronov et al. (2017) and Constantinescu et al. (2016) respectively, is it possible that they change their firing properties to represent any n-dimensional representational space, or they are bound, by evolution and/or any biological constraint, to reduce higher dimensional representations to the same format of the external physical environment? Future studies shall focus on this question by training subjects to parse and label multidimensional stimuli.

A wise use of functional neuroimaging methods and carefully designed behavioural paradigms will surely help answering these and many other questions.

Final remarks

I introduced the present work by asking three fundamental questions in the study of semantic representations. I can now provide short answers to these questions:

Question 1. How do semantic representations emerge in the human brain?

Semantic representations emerge in the human brain by means of the orchestrated plasticity of both memory and perceptual systems. The human brain is likely solving the symbol-grounding problem by locally modifying the representations of both symbols and their referent(s) so that they reflect the association with the complementary stimulus in the object-name association (the referent(s) and the symbols, respectively), at least for what I could witness being still at a relatively early stage of learning with my experiment. This process also involves the hippocampus, which might play a crucial role beyond episodic memory by recruiting spatial codes to support the construction of semantic representations as they were regions of a conceptual space.

Question 2. does the human brain recruit spatial codes for representing semantic spaces?

It seems so, even if the exact computations going on in different brain regions where I found these signals (right hippocampus, mPFC, EHC) are likely to be diverse and to serve different scopes.

Question 3. Does symbolic categorization facilitate generalization?

Yes, and this might be one of the key advantages of using symbols. I raised the possibility that this function might be linked to the hippocampal formation and related structures (such as mPFC) and their recruitment of spatial codes for representing semantic spaces.

REFERENCES

Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proceedings of the National Academy of Sciences 108: 4182–4187.

Althaus N, Westermann G. (2016) Labels constructively shape object categories in 10-month-old infants. Journal of Experimental Child Psychology 151: 5–17.

Althaus, N. & Plunkett, K., (2015). Categorization in infancy: Labeling induces a persisting focus on commonalities. Developmental Science, pp.1–11.

Aronov, D., Nevers, R., Tank, D.W. (2017) Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. Nature, 543, 719–722

Bao S. (2015) Perceptual learning in the developing auditory cortex. European Journal of Neuroscience 41: 718–724.

Behrens, T.E.J. et al. (2018) What is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron 100, 490-509

Bellgowan PSF, Bandettini PA, Van Gelderen P, Martin A, Bodurka J (2006) Improved BOLD detection in the medial temporal region using parallel imaging and voxel volume reduction. NeuroImage 25: 1244–1251.

Bellmund, J.L.S., Deuker, L., Navarro Schröder, T., Doeller, C.F., (2016) Grid-cell representations in mental simulation. eLife 5, e17089

Bellmund, J.L.S., Gardenfors, P., Doeller, C.F. (2018) Navigating cognition: Spatial codes for human thinking. Science. 362, 6415

Biederman, I. & Shiffrar, M., (1987) Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task.

Binder JR, Desai RH (2011) The neurobiology of semantic memory. Trends in Cognitive Sciences 15: 527–536

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19: 2767–2796.

Blundo, C., Ricci, M. & Miller, L., (2006). Category-specific knowledge deficit for animals in a patient with herpes simplex encephalitis. Cognitive Neuropsychology, 23(8), pp.1248–1268.

Bonnici XHM, Richter FR, Yazar Y, Simons JS (2016) Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. Journal of Neuroscience 36: 5462–5471.

Borghesani V, Pedregosa F, Buiatti M, Amadon A, Eger E, Piazza M (2016) Word meaning in the ventral visual path: a perceptual to conceptual gradient of semantic codice. Neuroimage 143: 128-140

Borghesani, V., Piazza, M. (2017) The neuro-cognitive representations of symbols: the case of concrete words. Neuropsych. 105, 4-17

Bornstein, M.H. & Arterberry, M.E., (2010) The development of object categorization in young children: hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. Developmental psychology, 46(2), pp.350–365.

Brants M, Bulthé J, Daniels N, Wagemans J, Op de Beeck HP (2016) How learning might strengthen existing visual object representations in human object-selective cortex. NeuroImage 127: 74–85.

Braunitzer, G. et al., (2017) The development of acquired equivalence from childhood to adulthood—A cross-sectional study of 265 subjects E. Ito, ed. PLoS ONE, 12(6), p.e0179525.

Brodt S, Gais S, Beck J, Erb M, Scheffler K, Schonauer M (2018) Fast track to the neocortex: A memory engram in the posterior parietal cortex. Science, 362: 1045-1048

Burgess N, Maguire EA, O'Keefe J (2002) The human hippocampus and spatial and episodic memory. Neuron 35: 625-641

Bush, D., Barry, C., Manson, D., Burgess, N. (2015) Using Grid Cells for Navigation. Neuron 87, 507–520

Capitani, E. et al., (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cognitive neuropsychology, 20(3), pp.213–261.

Caramazza, A. & Mahon, B.Z., (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. Trends in Cognitive Sciences, 7(8), pp.354–361

Clark IA, Kim M, Maguire EA (2018) Verbal Paired Associates and the hippocampus : the role of scenes, Journal of Cognitive Neuroscience 31:1-25

Cohen, L. et al., (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain : a journal of neurology, 123 Pt 2, pp.291–307.

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu Y, Abdi H, Haxby JV (2012) The representation of biological classes in the human brain. Journal of Neuroscience 32: 2608-2618

Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.J. (2016) Organizing conceptual knowledge in humans with a gridlike code. Science 352, 1464–1468

Dehaene S, Cohen L (2011) The unique role of the visual word form area in reading. Trends in Cognitive Science 15: 254-262

Dehaene, S., Cohen, L. (2007) Cultural Recycling of Cortical Maps. Neuron 56, 384-398

Desimone R, Schein SJ (1987) Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. Journal of Neurophysiology 57:835-868

DiCarlo, J.J., Zoccolan, D. & Rust, N.C., (2012) How does the brain solve visual object recognition? Neuron, 73(3), pp.415–434.

Doeller, C.F., Barry, C., Burgess, N. (2010) Evidence for grid cells in a human memory network. Nature 463, 657–661

Ellis, A.W., Young, A.W. & Critchley, E.M., (1989) Loss of memory for people following temporal lobe damage. Brain : a journal of neurology, 112 (Pt 6, pp.1469–1483.

Ezzati A, Katz MJ, Zammit AR, Lipton ML, Zimmerman ME, Sliwinski MJ, Lipton RB (2017) Differential association of left and right hippocampal volumes with verbal episodic and spatial memory in older adults. Neuropsychologia 93: 380-385

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. Journal of Neuroscience 33: 10552–10558.

Folstein JR, Palmeri TJ, Gauthier I (2013) Category learning increases discriminability of relevant object dimensions in visual cortex. Cerebral Cortex 23: 814–823.

Folstein, J. et al., (2015) Category Learning Stretches Neural Representations in Visual Cortex. Current directions in psychological science, 24(1), pp.17–23.

Friston KJ (2011) Functional and effective connectivity: a review. Brain Connectivity 1: 13–36.

Gainotti, G., (2012) Brain structures playing a crucial role in the representation of tools in humans and non-human primates. The Behavioral and brain sciences, 35(4), pp.224–225.

Garvert MM, Dolan RJ, Behrens TEJ (2017) A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. eLife 6:1–20.

Gibson EJ, Walk RD (1956) The effect of prolonged exposure to visually presented patterns on learning to discriminate them. Journal of Comparative and Physiological Psychology 49:239-242

Gibson JJ, Gibson EJ (1955) Perceptual learning: differentiation or enrichment? Psychological Review 62: 32-41

Giovanello KS, Keane MM (2003) Disproportionate deficit in associative recognition relative to item recognition in global amnesia. Cognitive Affective Behavioural Neuroscience 3: 186–194.

Goldstone RL (1994) Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General 123:178-200

Goldstone RL (1994) The role of similarity in categorization : providing a groundwork, Cognition 52:125-57.

Gopnik, A. & Meltzoff, A., (1987) The Development of Categorization in the Second Year and Its Relation to Other Cognitive and Linguistic Developments. Child development, 58(6), pp.1523–1531.

Hafting, T., Fyhn, M., Molden, S., Moser, M.B., Moser, E.I. (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801-806

Harnad S (1987) Categorical perception: the groundwork of cognition. New York: Cambridge University Press.

Hart, J., Berndt, R.S. & Caramazza, A., (1985) Category-specific naming deficit following cerebral infarction. Nature, 316(6027), pp.439–40.

Haxby J, Gobbidini IM, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293: 2425-2430

Heider, E.R. & Olivier, D.C., (1972) The structure of the color space in naming and memory for two languages. Cognitive Psychology, 3(2), pp.337–354.

Ison MJ, Quiroga RQ, Fried I (2015) Rapid encoding of new memories by individual neurons in the human brain article rapid encoding of new memories by individual neurons in the human brain. Neuron 87: 220–230.

Jacobs, J., et al., (2013) Direct recordings of grid-like neuronal activity in human spatial navigation. Nat. Neurosci. 16, 1188–1190

James, W., (1890) The principles of psychology. Available at: http://content.apa.org/books/10538-000.

Julian, J.B., Keinath, A.T., Frazzetta, G., Epstein, R.A. (2018) Human entorhinal cortex represents visual space using a boundary-anchored grid. Nat. Neurosci. 21, 191–194.

Kourtzi Z, Betts LR, Sarkheil P, Welchman AE (2005) Distributed neural plasticity for shape learning in the human visual cortex. Plos Biology 3:e204

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2:1–28.

Lawrence DH (1949) Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus. Journal of Experimental Psychology 39: 770-784

Lawrence Elbaum Associates. Burns, E.M. & Ward, W.D., (1978). Categorical perception--phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. The Journal of the Acoustical Society of America, 63(2), pp.456–468.

Lemus L, Hernàndez A, Luna R, Zainos A, Romo R, (2010) Do sensory cortices process more than one sensory modality during perceptual judgments ? Neuron 67:335-48.

Ley A, Vroomen J, Hausfeld L, Valente G, De Weerd P, Formisano E (2012) Learning of new sound categories shapes neural response patterns in human auditory cortex. Journal of Neuroscience 32: 13273–13280.

Ley A, Vroomen J, Formisano E, Brechmann A (2014) How learning to abstract shapes neural sound representations. Frontiers in Neuroscience 8: 1–11.

Liberman, A.M. et al., (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of experimental psychology, 54(5), pp.358–368.

Livingston, K.R., Andrews, J.K. & Harnad, S., 1998. Categorical perception effects induced by category learning. Journal of experimental psychology. Learning, memory, and cognition, 24(3), pp.732–753.
Livingstone M, Hubel D (1988) Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science, 240:740-749

Lupyan G (2015) The centrality of language in human cognition. Language Learning 66: 516:553

Lupyan G, Rakison DH, McClelland JL (2007) Language is not just for talking: redundant labels facilitate learning of novel categories. Psychological Science 18: 1077–1083.

Lupyan G, Thompson-Schill SL (2012) The evocative power of words: activation of concepts by verbal and nonverbal means. Journal of Experimental Psychology: General 141, 170-86

Mack ML, Love BC, Preston AL (2018) Building concepts one episode at a time: the hippocampus and concept formation. Neuroscience Letters 680, 31-38

Mahon, B.Z. & Caramazza, A., (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology Paris, 102(1–3), pp.59–70.

Mahon, B.Z. & Caramazza, A., (2009). Concepts and Categories: A Cognitive Neuropsychological Perspective. Annual Review of Psychology, 60(1), pp.27–51.

Mahon, B.Z. & Caramazza, A., (2010). Concepts and Categories: A Cognitive Neuropsychological Perspective. , pp.27–51.

Mahon, B.Z. & Caramazza, A., (2011). What drives the organization of object knowledge in the brain? Trends in Cognitive Sciences, 15(3), pp.97– 103.

McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review 102: 419-457

Miceli, G. et al., (2000). Selective deficit for people names following left temporal damage. Cognitive Neuropsychology, 17(6), pp.489–516.

Miller J, Watrous AJ, Tsitsiklis M, Lee SA, Sheth SA, Schevon CA, Smith EH, et al. (2018) Lateralized hippocampal oscillations underlie distinct aspects of human spatial memory and navigation. Nature Communications 9: 2423

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in Neurosciences 6: 414–417.

Moser, E.I., Moser, M.B., McNaughton, B.L. (2017) Spatial representation in the hippocampal formation: a history. Nat. Neurosci. 20, 1448-1464

Nau, M., Navarro Schröder, T. , Bellmund, J.L.S., Doeller, C.F. (2018) Hexadirectional coding of visual space in human entorhinal cortex. Nat. Neurosci. 21, 188–190

Newell, F.N. & Bulthoff, H.H., (2002) Categorical perception of familiar objects. Cognition, 85(2), pp.113–143.

Nikbakht N, Tafreshiha A, Zoccolan D, Diamond ME (2018) Supralinear and supramodal integration of visual and tactile signals in rats: psychophysics and neuronal mechanisms. Neuron 97: 626-639

Norman, G.R. et al., (1992) The correlation of feature identification and category judgments in diagnostic radiology. Memory & cognition, 20(4), pp.344–355.

O'Keefe, J. & Dostrovsky, J. (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 34, 171-175

O'Keefe, J., Nadel, L. The Hippocampus as a Cognitive Map (Clarendon, 1978).

Oakes, L.M., Coppage, D.J. & Dingel, A., (1997) By land or by sea: the role of perceptual similarity in infants' categorization of animals. Developmental psychology, 33(3), pp.396–407.

Obleser J, Zimmermann J, Van Meter J, Rauschecker JP (2007) Multiple stages of auditory speech perception reflected in event-related fmri. Cerebral Cortex 17: 2251-2257

Olman CA, Davachi L, Inati S (2009) Distortion and signal loss in medial temporal lobe. Plos One 4: e8160

Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in MATLAB/GNU Octave. Frontiers in Neuroinformatics 10: 27

Op de Beeck H, Baker CI (2010) The neural basis of visual object learning. Trends in Cognitive Science 14:22-30

Op de Beeck H, Baker CI, DiCarlo JJ, Kanwisher N (2006) Discrimination training alters object representations in human extrastriate cortex. Journal of Neuroscience 26: 13025–13036.

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Thomas M, Leff AP (2010) Comparing families of dynamic causal models. Plos Computational Biology 6: e1000709

Peron, R.M. & Allen, G.L., (1988). Attempts to train novices for beer flavor discrimination: a matter of taste. The Journal of general psychology, 115(4), pp.403–418.

Piazza M, Pinel P, Le Bihan D, Dehaene S (2007) A magnitude code common to numerosities and number symbols in human intraparietal cortex. Neuron 53: 293-305

Qu, L.P. et al., (2016). De Novo Emergence of Odor Category Representations in the Human Brain. Journal of Neuroscience, 36(2), pp.468–478.

Quiroga RQ (2012) Concept cells: the building blocks of declarative memory functions. Nature Review Neuroscience 13: 587–597.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435: 1102–1107.

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of “ what ” and “ where ” in auditory cortex, Proceedings of the National Academy of Sciences 97:11800-11806

Roberson, D. et al., (2005) Color categories: Evidence for the cultural relativity hypothesis. Cognitive Psychology, 50(4), pp.378–411.

Robinson, C.W. et al., (2012) The Role of Words in Cognitive Tasks: What, When, and How? Frontiers in Psychology, 3, p.95.

Rodman, H.R., Skelly, J.P. & Gross, C.G., (1991) Stimulus selectivity and state dependence of activity in inferior temporal cortex of infant monkeys. Proceedings of the National Academy of Sciences , 88(17), pp.7572–7575.

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

Rosch, E., (1975) Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104(3)

Rumiati, R.I. & Foroni, F., (2016) We are what we eat: How food is represented in our mind/brain. Psychonomic Bulletin & Review, 23(4), pp.1043–1054

Sacchett, C. & Humphreys, G.W., (1992) Calling a squirrel a squirrel but a canoe a wigwam: a category-specific deficit for artefactual objects and body parts. Cognitive Neuropsychology, 9(1), pp.73–86.

Samson, D. & Pillon, A., (2003) A Case of Impaired Knowledge for Fruit and Vegetables. Cognitive Neuropsychology, 20(3–6), pp.373–400.

Schmidt CF, Degonda N, Luechinger R, Henke K, Boesiger P (2005) Sensitivity-encoded (SENSE) echo planar fMRI at 3T in the medial temporal lobe. NeuroImage 25: 625–641.

Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron 91(6): 1402-1412.

Schurger A, Pereira F, Treisman A, Cohen JD (2010) Reproducibility distinguishes conscious from nonconscious neural representations. Science 293(5539):2425-30.

Schurger A, Sarigiannidis I, Naccache L, Sitt JD, Dehaene S (2015) Cortical activity is more stable when sensory stimuli are consciously perceived. Proceedings of the National Academy of Sciences 112(: 2083-2092

Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesion. Journal of Neurology, Neurosurgery, and Psychiatry 20:11-21

Shallice, T. & Cooper, R.P., (2013) Is there a semantic system for abstract words? Frontiers in Human Neuroscience, 7, p.175.

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Adam M (2011) Commonality of neural representations of words and pictures. NeuroImage 54: 2418–2425.

Sigala, N. & Logothetis, N.K., (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature, 415(6869), pp.318–320.

Simanova I, Hagoort P, Oostenveld R, Van Gerven MAJ (2014) Modality-independent decoding of semantic information from the human brain. Cerebral Cortex 24: 426-34.

Spiers HJ, Maguire EA, Burgess N (2001) Hippocampal amnesia. Neurocase 7: 357–382.

Stalnaker TA, Cooch NK, Schoenbaum G (2015) What the orbitofrontal cortex does not do. Nature Neuroscience 18, 620-627.

Stemmler, M., Mathis, A., Hertz, A.V.M. (2015) Connecting multiple spatial scales to decode the population activity of grid cells. Science Advances 1, e1500816

Stephan KE, Penny WD, Danizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. NeuroImage 46: 1004-1017

Theves S, Fernandez G, Doeller CF (2019) The Hippocampus encodes distances in multidimensional feature space. Current Biology 29, 1-6

Tolman, E.C. (1948) Cognitive maps in rats and men. Psychol. Rev. 55, 189–208

Tsunada, J. & Cohen, Y.E., (2014) Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. Frontiers in Neuroscience, 8, p.161.

Vanderplas JM, Sanderson WA, Vanderplas JN (1964) Some task-related determinants of transfer in perceptual learning. Perceptual and Motor Skills 18: 71–80.

Waller TG (1970) Effect of irrelevant cues on discrimination acquisition and transfer in rats. Journal of Comparative and Physiological Psychology 73: 477 – 480

Warrington, E.K. & Mccarthy, R., (1983) Category specific access dysphasia. Brain, 106(4), pp.859–878.

Warrington, E.K. & Shallice, T., (1984) Category Specific Semantic Impairments. Brain, 107, pp.829–854.

Watson AB, Pelli DG (1987) QUEST: a bayesian adaptive psychometric method. Perception and Psychophysics 33: 113-120

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014) Orbitofrontal cortex as a cognitive map of task space. Neuron 81(2): 267-279.

Yazar Y, Bergstrom ZM, Simons JS (2017) Reduced multimodal integration of memory features following continuous theta burst stimulation of angular gyrus. Brain Stimulation 10: 624-629