Les concepts abstraits comme problème pour une théorie de la connaissance

Guido Löhr

PSL UNIVERSITÉ PARIS

ENS ÉCOLE NORMALE SUPÉRIEURE

RUHR UNIVERSITÄT BOCHUM RUB

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PSL

Préparée à [Etablissement de préparation]

Dans le cadre d'une cotutelle avec [Etablissement partenaire]

Les concepts abstraits comme problème pour une théorie de la connaissance

The Problem of Abstract Concepts for a Theory of Situated Cognition

Soutenue par

Guido LÖHR

Le 22.05.2020

Ecole doctorale n° 540

Lettres, Arts, Sciences humaines et sociales

Spécialité

Philosophie

Composition du jury :

Paul, EGRÉ Directeur de recherche au CNRS Institut Jean Nicod, UMR 8129	*Président*
Corinna, MIETH Professeur Ruhr Uni Bochum	*Rapporteur*
Matthias, Unterhuber Dr. Ruht Uni Bochum	*Rapporteur*
Albert, NEWEN Professeur Ruhr Uni Bochum	*Examinateur*
Joachim, Horvath Professeur Ruhr Uni Bochum	*Examinateur*
Markus, WERNING Professeur Ruhr Uni Bochum	*Directeur de thèse*
François RECANATI Proesseur au Collège de France	*Directeur de thèse*

Institut | Nicod

CNRS-EHESS-ENS

RÉSUMÉ

Cette thèse cumulative défend l'idée qu'une grande partie de la littérature interdisciplinaire traitant des concepts confond les concepts avec ce qui est utilisé pour appliquer les concepts. Plus précisément, cette thèse soutient que les questions relatives au contenu (ce sur quoi porte le concept, sa sémantique) ont été confondues avec les questions relatives à notre accès épistémique à ce contenu (ce que nous savons de ce contenu). Une fois cette distinction établie, il est possible de résoudre un certain nombre de problèmes qui ont contraint la littérature conceptuelle pendant des décennies. Premièrement, il devient alors possible de noter que les types de concepts auxquels les psychologues se sont intéressés pour expliquer le problème de l'application des concepts, comme dans la catégorisation, ne doivent pas nécessairement traiter des problèmes sémantiques de la compositionalité et de la systématicité. Deuxièmement, il devient également possible de laisser place à la possibilité empirique que des concepts abstraits, c'est-à-dire des concepts qui ne i s'appliquent pas à des objets physiques concrets avec lesquels nous avons un contact sensoriel direct, puissent être mieux expliqués par une approche de la cognition située ou empiriste, c'est-à- dire par des mouvements corporels, des représentations sensorimotrices, ou des représentations de situations et d'états introspectifs.

MOTS CLÉS

Concepts Abstraits, Contenu, Catégorisation, Situated Cognition, Compositionalité

ABSTRACT

The main argument of this interdisciplinary cumulative thesis is that a large part of the interdisciplinary literature on concepts conflates concepts with that which we apply our concepts in terms of. More precisely, I argue that it conflates questions of content (what a concept is about, i.e., its semantics) with questions of our epistemic access to this content (what we know about what our concepts are about). Once this distinction is understood we can solve a number of problems that have riddled the concept literature for decades. First, we can see that the kind of concepts that psychologists are interested in to explain the epistemic problem of how we apply our concepts, e.g., in categorization, need not address the semantic problems of compositionality and systematicity. Secondly, we can now make room for the empirical possibility that abstract concepts, i.e., concepts that do not apply to concrete physical objects we have direct sensory contact with, are best explained by a situated or empiricist approach to cognition, i.e., by means of bodily movements, sensorimotor representations, or representations of situations and introspective states.

KEYWORDS

Abstract Concepts, Content, Categorization, Situated Cognition, Compositionality

Das Wort "Begriff" wird verschieden gebraucht, teils in einem psychologischen, teils in einem logischen Sinne, teils vielleicht in einer unklaren Mischung von beiden. (...) Die Frage, ob dieses oder jenes zweckmaßiger sei, möchte ich als weniger wichtig beiseite lassen. Man wird sich leicht über die Ausdrucksweise verständigen, wenn man einmal anerkannt hat, daß etwas da ist, was eine besondere Benennung verdient. (...) Hieraus entspringen ja leicht Widerspruche, die nicht meiner Gebrauchsweise zur Last fallen.

The word 'concept' is used in various ways; its sense is sometimes psychological, sometimes, logical, and sometimes perhaps a confused mixture of both. (...) The question whether this or that use is more appropriate is one that I should like to leave on one side, as of minor importance. Agreement about the mode of expression will easily be reached when once it is recognized that there is something that deserves a special term. (...) This readily gives rise to contradictions, for which my usage is not to blame.

Frege (1892) from *Über Begriff und Gegenstand* (*On Concept and Object*, translated by Geach and Black, 1951)

Acknowledgements

I thank my supervisors Markus Werning and François Recanati for their philosophical input, support and feedback. I thank Peter Brössel for his guidance and advice. Most of the present work had previously been discussed with Dimitri Mollo, Juan Loaiza, Edouard Machery and the many helpful anonymous reviewers of the respective journals I sent my manuscripts to. Many of the chapters also greatly benefited from discussions with my research group in Bochum and Osnabrück on Situated Cognition (especially Beate Krickel, Albert Newen, Julia Wolf, Elmarie Venters, Julian Packheiser, Matej Kohár, Samantha Eli, Benjamin Angerer etc.) as well as the PhD students at the Institut Jean Nicod in Paris (especially Romain Bourdoncle, Luca Gaspari and Marco Inchingolo). The introduction, conclusion and chapter 4 benefited from discussions with Steffen Koch. I also want to thank Ellen Fridland who patiently read very early drafts of chapter 2 without discouraging me as well as my former supervisors Richard Moore, Michael Pauen and Geert Keil who commented on material in chapters 1 and 2. Chapter 4 benefited a lot from discussions with Esa Diaz Leon and Sarah Sawyer. I thank the German Research Council and the Research School Plus in Bochum for their financial support. All of the material was previously presented at various conferences, so I would like to thank the organizers and participants especially of the conferences hosted by the European Society of Philosophy and Psychology (ESPP) and the European Network of Social Ontology (ENSO). Most of all, I am grateful to my parents, Jutta and Peter, for their unconditional support.

Publications used in this thesis:

Chapter 1: Löhr, G. (2018, published online). Concepts and categorization: do philosophers and psychologists theorize about different things? *Synthese*.

https://link.springer.com/article/10.1007%2Fs11229-018-1798-4

Chapter 2: Löhr, G. (2017). Abstract concepts, compositionality, and the contextualism-invariantism debate. *Philosophical Psychology*, 30(6), 689-710.

https://www.tandfonline.com/doi/abs/10.1080/09515089.2017.1296941?journalCode=cphp20

Chapter 3: Löhr, G. (2019). Embodied cognition and abstract concepts: Do concept empiricists leave anything out? *Philosophical Psychology,* 32(2), 161–185.

https://www.tandfonline.com/doi/abs/10.1080/09515089.2018.1517207?journalCode=cphp20

Chapter 4: Löhr, G. (2019, published online). Social constructionism, concept acquisition and the mismatch problem. *Synthese*.

https://link.springer.com/article/10.1007/s11229-019-02237-2

Conventions

Titles of books and articles, emphasis and some Latin terms	Italic letters	*Philosophical Investigations* *as, prima facie*
Technical and unusual terms, quotes	Double quotation marks	“concept”, “mansplaining” “Cogito ergo sum”
Lexical expressions	Simple quotation marks	‘electron’
Names of concepts	All capital letters	ELECTRON

TABLE OF CONTENTS

Introduction

1 What are concepts

This is a cumulative dissertation. This means that each of the chapters below is written as a single stand-alone journal article that has been published in a peer-reviewed journal, such as *Philosophical Psychology* or *Synthese.* Still, each paper contributes to a common interdisciplinary goal. This goal is a theory of the nature and psychology of concepts, in particular abstract concepts like TRUTH, ART, KNOWLEDGE, or DEMOCRACY.

The concept of concept is among the most fundamental we have in cognitive science, especially in psychology, linguistics and philosophy. Concepts are thought to underlie and thus partly explain how we can think or speak about the world and some have argued that concepts are even essential to early perception, i.e., the way we see the world. The concept of concept is so fundamental to our investigation of the mind that it has occupied both philosophers and cognitive scientists at least since Plato to understand what concepts are and how they impact our access to the world.

At conferences, I have often heard that the notion of concept, due to its long history and its interdisciplinary use, has become too "messy" to be useful or that nobody really knows what we mean by the term anymore, suggesting a very dismissive and pessimistic attitude towards both the notion itself and research on this notion. It implies that engaging in the convoluted literature on what concepts are is doomed from the beginning simply because nobody in the debate really knows what they are talking about, or at least whether they are talking about the same thing. For this reason, many philosophers, linguists and psychologists nowadays try to avoid using the notion, or at least avoid the attempt to define it.[1]

Another common approach in today's cognitive science community is to side-step the difficulty of defining what we mean by the term 'concept' and to introduce the term by means of the idea of a family resemblance. The idea of a family resemblance is inspired by an observation by Ludwig Wittgenstein (1953) in his *Philosophical Investigations* that many ordinary language terms denote entities that have interestingly overlapping but no single set of necessary and sufficient commonalities. Some have interpreted Wittgenstein as suggesting that this shows that

[1] This claim is based on observations I have made at several cognitive science and concept conferences in the last years. It is not based on published work. However, I hope that my observation is shared by the reader assuming they are part of this particular community.

nothing more can or need to be said to characterize a concept except simply listing features or examples that we take to be commonly associated with it.

The idea that the term 'concept' cannot and need not be defined and that it may be best understood in terms of some kind of a family resemblance relation cannot be right. First, 'concept', at least as it is used in psychology and philosophy, is a technical term with a lot of theoretical burden and motivation. It is a key notion that underlies much experimental research that we need to make sense of if we want to understand these experiments and the theories that motivate them. Technical terms, unlike ordinary language terms, are invented *by us* because they fulfill a certain theoretical or explanatory role. We have to be able to clearly state what this role is supposed to be for it to do real explanatory work. Since we invented these concepts for a certain purpose we should be able to do so in a relatively clear manner.

There are many terms that arguably work very differently. We may think of early research on the essential properties of 'water'. Here it makes sense to say that researchers did not have to clearly define what they mean by 'water' in order to successfully study it. Water seems to be just the stuff we drink, swim in and that falls from the sky when it rains. Wanting to know what this *stuff* essentially is and what explains its contingent properties (e.g., why it boils), does not require much more than an approximate description of the properties that reliably correlate with it. Similarly, ordinary language terms like 'game' seem to be impossible to define and since this term does not motivate or underlie much experimental work in psychology (except maybe in game theory where the notion is however used in a more specific sense) there is usually no need for a precise definition.

The concept of concept in psychology, linguistics and philosophy is different. It plays a crucial theoretical role and we must know what we are talking about when we base our theories and experiments on a technical notion. It is neither an ordinary language term, nor a term that picks out a certain substance in every possible world. Instead, the meaning of the term 'concept', as it is used in philosophy and psychology, is described functionally. Functionally described terms are names for whatever serves a certain function or plays a certain theoretical role that we identify as important or even crucial, e.g., when reflecting about our mental life. The term 'concept' is thus more like the concept of photosynthesis than the concepts of water or game. Nobody would argue that biologists who posit a process of photosynthesis to explain how plants get their energy do not need to know what exactly they mean by this term.

So, what is this crucial theoretical role that 'concept' ought to play? I argue that, most minimally, a concept is that which allows us *to think* about something *as something*. It is that which we think in terms of. It is what is applied to objects *in thought*, i.e., when believing, desiring, hoping, guessing, and so forth. For example, if you believe that you are reading a dissertation you are thinking of the text in front of you *as a dissertation*, i.e., you are correctly applying your concept of dissertation to this text, *in thought*.

Since any application of concepts to objects can, *essentially*, be true or false, i.e., it has correctness conditions or conditions of satisfaction, concepts are a kind of representation. To say that concepts are representations does not result in any ontological commitment. I have neither said that representations exist nor what kind of entity actually plays the role of a representation. This is because the notion of representation, too, is functionally defined – they are those things that have correctness conditions. Thus, I have *not* argued that I take concepts or representations to pick out, for example, mental entities. I take it that there may be all kinds of non-mental representations (paintings for instance). My descriptions so far also do not commit me to the idea that what is picked out is a natural or a scientific kind, i.e., a kind that exists independently of our interests. It may turn out that 'representation', 'concept' and 'belief' do not pick out anything real or mind-independent in the world.

However, concepts are supposed to be more than mere representations. They are supposed to be a special kind of representation. What differentiates them from other representations? They certainly have a lot in common with representations in general besides having correctness conditions. First, both conceptual and non-conceptual representations ought to be *stable*. Think of a low-level vision, say a set of neurons that only respond to a certain orientation, color or shape. Assuming for the sake of argument that they are in fact a kind of representation, such neurons respond correctly to, say, a vertically orientated object. In order to be able to apply the same representation correctly to different objects in different circumstances we would like our orientation detector representations to reliably detect the same orientation correctly in various situations. It would not be very helpful to our visual system if it detected in one situation *correctly* a vertical line and in another *correctly* a horizontal line. Instead, our representation of vertical lines *essentially* correctly responds only to vertical lines and incorrectly to horizontal lines. Stability is also an essential property of conceptual representations. The concept of dog is *essentially* about dogs. If it were sometimes about dogs and at other times about cats it would cease to be the concept of dog and would now be the concept DOG OR CAT. If it applied correctly only to typical cats then it would be the concept TYPICAL CAT.

Secondly, both conceptual and non-conceptual representations apply not only to different objects, but we may ascribe different conceptual and non-conceptual representations to the same object. Just as the same car has a number of different shapes, orientations and functions, so is Donald Trump not just a human being, but a man, a president, a business man, a husband and so forth. Thus, the same representation and the same concept can apply to a number of objects, and the same object can be ascribed a number of representations and concepts. So, again, what distinguishes conceptual from non-conceptual representations if they have all the same properties? I argue that the main difference between conceptual and non-conceptual representations is that only the former feature in *thoughts*, i.e., beliefs and desires. So, to fully differentiate conceptual from non-conceptual representations we need to consider not just properties of representations but properties of thought.

2 Conceptual and non-conceptual representations

For an intuitive way to approach the problem of distinguishing conceptual from non-conceptual representations consider the Müller-Lyer illusion. This visual illusion can be triggered by means of three stylized arrows that are of the same length except that the endpoints of the arrow consist of arrow heads that either point inwards or outwards. Depending on whether the arrow heads point inwards or outwards, the shaft of the arrow appears either longer or shorter (see Figure 1). Interestingly, when people are asked to interact with a physical instantiation of the lines, e.g., if they are asked to grasp it, their motor response is not affected by the visual illusion (e.g., Goodale and Humphrey, 1998, Bruno, 2001). This is an incredibly interesting and counter-intuitive fact that clearly illustrates the need for two kinds of representations, one of them is potentially conscious and part of our beliefs and the other is not. We do not usually believe that the length is the same (at least not based on our perception alone), but we nonetheless represent it in some way as being of the same length. We just do not represent it as being of the same length *in thought*.

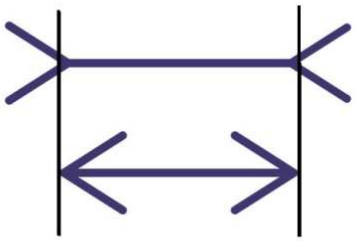

Figure 1: Image that exhibits the Müller-Lyer illusion.

So, when looking at the three arrows, certain perceptual representations represent the length of the shafts of the arrows as being of the same length while other perceptual representations present them as being of a different length. Only the latter are the representations that usually justify our (false) belief about the lengths of the arrows. Since the former do not and cannot feature in our beliefs, i.e., they are not and cannot be applied to objects *in thought,* they cannot, therefore, by definition, be conceptual representations. Thus, the Müller-Lyer illusion suggests that we apply two contradictory kinds of representation to the same object. I argue that one perceptual representation is non-conceptual, non-conscious and important for immediate actions, while the other perceptual representation is conceptual, potentially conscious and crucial for more epistemic purposes, e.g., to justify our beliefs about the world.

Thus, only the perceptual representations that give rise to our belief that the lines are of different lengths are conceptual. This is not necessarily so because it is vastly different in nature when compared with the non-conceptual motor representation, but simply because it is the one that features in our thoughts about the object. We *believe* that the object has lines with different lengths even though early visual non-conceptual representations represent it to the motor system correctly as being of the same length. Of course, we can use some external representations to measure the length and then form the belief that the object is in fact of the same length. However, the application of the concept of same length to the lines is not justified by means of our internal perceptual representations. Instead, it is justified by us *believing* that a certain external object is a more accurate representation of the length of the given object than our internal consciously available resources.

So, for now, the most important thing we need to know about the meaning of 'concept' is just that it ought to pick out a kind of representation that we apply to objects *in thought*. By 'thought'

we mean potentially conscious personal level epistemic states (they are attributed to a person and not a brain for example) like beliefs and desires. There may be other notions of concept in the literature (some philosophers like Dummett, 1993 or Kenny, 2019 use the term 'concept' to pick out certain kinds of abilities), but these other notions may just pick out different interesting phenomena. In this thesis, I am interested in the nature and application of those kinds of representations we use *in thought*, which I call 'concept'.

3 Requirements for a theory of concepts

There are other properties of thought that further allow us to distinguish conceptual from non-conceptual representations and that constrain our theories of concepts. One of the most interesting properties of thought is its immense productivity. Productivity is often defined in terms of the notion of infinity (we can in principle have an infinite number of different thoughts). However, more common-sense-friendly introductions are even more powerful to make the distinction between concepts and other kinds of representations apparent. The notion of productivity I find most interesting for the purpose of this thesis is the following non-binary common-sense notion from economics. Imagine two factories that both turn tree trunks into tables. Both factories are productive. They produce tables. However, the first factory produces five tables per tree trunk while the second factory produces only one table per tree. The second factory is therefore less productive than the first factory.

Translated to the question of productivity of thought, it is striking that we can obtain many more conceptual representations compared to the conceptual representations we actually store, simply by combining them. Using certain syntactic recursive operations, we could in principle even generate an infinite number of different thoughts which, in principle, i.e., if our long-term and short-term memory were large enough, we could even comprehend. For example, we cannot just comprehend the sentence "John is the father of Jenny", but the sentence "John is the father of Jenny who is the mother of Bob who is the dancer of the group that first started the war, which led to a number of new inventions…" and so forth. There is no reason to think that we had this very complex representation already stored in our long-term memory. Instead, it is more plausible that we were able to generate it "on the fly" by combining representations we do store.

More interesting than the possibility of generating a potentially infinite string of concepts is that the productivity of thought allows us to generate all kinds of new thoughts that we would

never even have dreamed of forming. This allows us to think and talk about all kinds of new things even if we lack words for them, which is crucial for a species that relies strongly on the ability to adapt to previously unexperienced circumstances. For example, we can think that the house on the planet Mars that is painted green will never look as good as the sun setting on an open field. We can think that love is more valuable than a funny game played on a sad Sunday morning. But we can also think that this creature there with the long teeth might become dangerous to our community or that the stuff in the atmosphere called CO2 might destroy the environment. Some of those are absurd thoughts that I have never entertained previously in my life and that I have never heard expressed by anyone. Still, we can think and even share these thoughts even if we have little understanding of what we would have to do if we wanted to investigate whether they are true. Thus, thought and language are incredibly productive systems and this productivity requires a certain kind of representation. In other words, this productivity puts interesting constraints on our notion of concept that further helps to distinguish concepts from other representations.

The most crucial feature we require conceptual representations to possess if we want them to explain the productivity of thought besides, again, stability, is that they have to compose *in a certain way*. Note that not all representations compose in a way that gives us something that is new over and above what we put into the compositional process. Paintings in a museum do not compose in this sense. The Mona Lisa combined with Duchamp's *Fountain* does not give us anything else but two paintings located spatially next to each other. This combinatory operation is not very productive and can be compared to a factory that takes two trunks and simply glues them together. It is not clear whether the customer who ordered a table from the factory would be very pleased with the delivery of two tree trunks glued together. Similarly, representations that allow us to grasp a cup combined with representations that allow us to laugh may not give us anything but the conjunction of both. Conceptual representations however need to combine in a way that generates more than the conjunction both. It ought to generate new complete truth evaluable thoughts that we may have never heard expressed by anyone and that we may never have entertained before.

Furthermore, concepts ought to combine in highly systematic ways. It cannot be the case that we can think that John borrowed a glass of wine from Sally, but not that a glass of wine borrowed Sally from John. Again, certainly this is a very strange thought that we are probably not able to properly visualize or even make perfect sense of because we do not generally view glasses as agents that could borrow people. Still, we are nonetheless able to think it. The

sentence is true if and only if a glass of wine borrowed Sally from John. For example, imagine a world where glasses developed brains and arms to carry people. If we suddenly woke up in such a strange world, this is the thought we would then want to form to be able to behave appropriately in this new situation. A system that did not allow us to generate such a thought would be highly deficient. The system we actually have is much more powerful.

That we cannot visualize or really understand what it could mean that a glass borrowed a person is not necessarily good evidence that we are not able to think it. There are all kinds of thoughts we can think and that have truth conditions without us really understanding them in the sense of being able to verify whether they are true or false. Consider the following sentence: "each of the molecules of water contains two oxygen and three hydrogen atoms, connected by covalent bonds". There is a sense in which most lay people are not able to really imagine or visualize what it would mean for this sentence to be true. The sentence uses rather abstract scientific terminology that many people are not very familiar with. Still, if a lay person were to say that the sentence is true, this person would actually be wrong. The sentence and the thought it expresses are in fact false. So, did the lay person uttering the sentence say something false or did they say nothing at all? Many philosophers share the intuition that the speaker said and thus also thought something wrong without there being much understanding of what exactly the sentence referred to. So, many philosophers today assume that we should not base our theory of thought on our intuitions on what we are able to really understand and instead base it first of all on the question of what it takes to relate to an object in thought.

Finally, to explain how concepts can generate new complete thoughts they need to compose in ways that cannot further require any adjustment by context. So, concepts need not only be stable, but they need to compose systematically in a very strict way that is guided merely by a finite set of syntax rules. The main motivation for strong requirements of stability, compositionality and systematicity, i.e., as being context-independent, is that we have a theoretical need for terminology and a conceptual framework to talk about the most fundamental unit of thought. In order for concepts to play this role, they cannot further be disambiguated by a more fundamental level of thought. If we needed a more fundamental level of thought that could make sense of certain ambiguities in thought then we would not be talking about the most fundamental level of thought, i.e., we would not be talking about concepts. This is why thought and concepts have to be systematic and compositional in a *very strong sense* that cannot be disambiguated further by means of world knowledge or discourse context. If

concepts are supposed to be the most fundamental level of thought, then there cannot be a more fundamental level in terms of which we could disambiguate concepts and thoughts.

Thought then contrasts with natural language, which *is* incredibly ambiguous, unsystematic and arguably not compositional in the strong sense that we require for concept compositionality (Recanati, 2010). This ambiguity however can easily be resolved by means of a more fundamental level of representation in terms of which we can formulate the ambiguity. Again, this is the level of thought. The relation between language and thought is easiest to see for names. The sentence "Alex is happy" is ambiguous if you know two different people who are called 'Alex'. We can make sense of the idea that we have to do with two different sentence types expressing two different propositions if we assume that we have two concepts of Alex, one referring to Alex1 and one referring to Alex2. Similarly, the ambiguity of the sentence "let's go to the bank" can be resolved by means of translating this level of representation to a more fundamental conceptual level LET'S GO TO THE FINANCIAL INSTITUTION versus LET'S GO TO THE RIVERBANK. This means that language does not *have* to be compositional and systematic in this very strong sense if we have a more powerful level of representation that we can translate or disambiguate linguistic representations in terms of.

Imagine concepts were not systematic and compositional in the strong sense that the meaning of the whole is derived merely from the meaning of its parts and a finite set of rules of combination. Take the English sentence "John is a fake friend" and compare it to "John has a fake mullet". In a compositional semantics, we may ask whether the linguistic sign 'fake' corresponds to the same concept in both sentences. Assume for a moment that 'fake' contributes a different meaning to the first sentence than the second. For example, we may argue that 'fake' in the first sentence has a function that gives us the complement set of its argument 'friend'. Since fake mullets however are still mullets (a certain kind of wig for example), 'fake' in the second sentence cannot have this function. This suggests that we have to do with a linguistic ambiguity and 'fake' is associated with two different concepts. Since we needed concepts to make this supposed ambiguity and lack of systematicity and compositionlity of the word 'fake' explicit, we cannot further assume that concepts themselves are ambiguous in this sense. So, again, while linguistic representations may not be systematic and compositional, conceptual representations, by means of their theoretical role, *must* be. If an alleged theory of concepts cannot explain how concepts can be stable (context-insensitive) and compositional in a strongly systematic way that can explain productivity of thought and disambiguation of sentences in natural language, then it is not a theory of concepts. It is not a theory of the most fundamental

unit of thought, even if it may still be an interesting or even true theory of another kind of representation.

4 "Theories of concepts"

Over the past decades there have been a number of theories proposed that run under the label "theory of concept". However, it remains controversial whether many of these theories can in fact meet the conditions for a successful theory of concept specified above (context-independent compositionality, context-independent systematicity and stable content). In fact, I shall argue in chapter 1 that many of the theories of concepts proposed especially in psychology are not even designed to give a theory of concepts as the notion of concept has been introduced above and how it is usually used in philosophy of language and mind. Instead, I argue below that most of these theories are better understood as theories of what I call "categorization devices", i.e., devices that explain not what concepts are but how we apply them, e.g., in categorization or when making inferences. I will discuss how those two very different kinds of entity (concepts and categorization devices) relate to each other in the first chapter and the conclusion.

However, for now, I will simply summarize the different contenders for a successful theory of concept without any judgment as to whether they are in fact successful theories of concepts or even intended as such. The reader will notice that I will summarize the different theories and objections to them in a rather quick manner. The reasons for this are that, first, a more detailed description of the different theories is not needed for the purpose of this thesis and can be found in various monographs on these theories, in particular Murphy (2002) or Machery (2009). Secondly, the different objections have been widely discussed in the philosophical and psychological literature in the past decades and are familiar to many philosophers and psychologists interested in this topic. Most importantly, I do not discuss especially the objections and possible replies in detail because I aim to resolve them in the first chapter by means of my distinction between concepts and categorization devices. In fact, I do not think for example that compositionality is a problem for prototype theory if we consider it a theory of categorization devices as opposed to a theory of concepts. Similarly, I do not think that most empirical objections to definitions pose a problem for definitionism if we consider it purely a theory of some word meanings and not a theory of categorization or concepts in general.

4.1 The classical theory

The theory that all the other theories introduced here are a response to is called *the classical view* or *definitionism*. It is called the classical view because it has been the standard view of concepts for most of the history of Western Philosophy (Laurence & Margolis, 1999; Prinz, 2002). According to definitionism, most lexical concepts, i.e., concepts that have a word in ordinary natural language, are complex conceptual representations structured in terms of necessary and sufficient features or conditions that an entity must possess or meet in order to belong to the category that the concept represents. For example, it is assumed that the expression 'bachelor' is associated with the complex concept UNMARRIED ADULT MALE, such that in order to apply the word 'bachelor' correctly to a person, we must also be able to correctly apply the concepts of unmarried adult male to them.

It is at this point important to emphasize that definitionism is more a theory about language or linguistic meaning than about concepts. Again, it states that most lexicalized concepts are like definitions. It is not controversial that complex concepts and phrases are definitional. Thus, the linguistic phrase "apple that has red seeds" may arguably pick out the complex concept APPLE THAT HAS RED SEEDS, which specifies necessary and sufficient conditions for its application. The phrase as well as the complex concept only apply correctly to apples that have red seeds. In other words, for something to be an apple with red seeds we also need to be able to apply the concepts of red, apple and seed to it. Thus, to be absolutely explicit, what is controversial is whether most of our words are associated with complex *definitional* concepts or not and *not* whether concepts are definitional (see, Margolis & Laurence, 2019 for this confusion).

The most pressing empirical reason against definitionism is the lack of successful attempts to define ordinary language terms (Laurence & Margolis, 1999; Prinz, 2002). Again, according to a classic example by Wittgenstein (1953) we cannot even define the everyday word 'game'. Wittgenstein famously argued that no single set of necessary and jointly sufficient conditions can capture all instances that we intuitively identify as games. Imagine for instance that games are defined in terms of 'is fun' and 'can be played by more than two players'. How then does one correctly classify the game Solitaire? Solitaire is a game even though many people find it relatively boring and even though it is usually played by only one person. Moreover, would it be more clearly a game if people did not find it boring? Could we also not imagine playing solitaire together taking turns? Put differently, it seems that no single rule can prescribe all correct applications of a concept, suggesting that definitions are neither what captures the meaning, nor the application conditions of many ordinary language words.

Perhaps, one may argue that the term ‘game’ is exceptionally difficult to define and that, surely, some ordinary language words *are* definable. However, even classic examples of words that do seem to be construable in terms of necessary and sufficient conditions are problematic. It is difficult to find an uncontroversial set of necessary and jointly sufficient conditions even for the word 'bachelor'. The word 'bachelor' could just mean *unmarried male*, but this definition does not explain why, for example, we do not classify young boys as bachelors. Perhaps ‘bachelor’ just means *unmarried adult male*. Still, most people would not classify the pope as a bachelor. At least it would be rather controversial to do so (Prinz, 2002). Hence, there seems to be more to our conceptual abilities than is captured by necessary and sufficient conditions.

Perhaps definitions of simple (as opposed to composed or complex) ordinary words are hard to find because they are implicit. They just have not been explicated yet (Rey, 1983). But then how can we ever think to falsify definitionism? And how can it ever explain our conceptual abilities, such as inference-making or categorization, if the definitions cannot be brought to consciousness? Note that in order for a definition to allow us to appropriately apply a concept, all its application conditions ought to be immediately and simultaneously available. But then, why would they be so difficult to find?

A related and very influential argument against definitions goes back to Saul Kripke’s (1972) and Hilary Putnam’s (1975) work on proper names and natural kind terms that will be important throughout the thesis (especially chapters 1, 3 and 4). Kripke and Putnam pointed out that we can make true statements about an entity even if we are wrong or ignorant about its essential features, i.e., even if we do not know the proper application conditions of a concept. For example, most people are ignorant about the essential features of gold, but can still easily recognize gold reliably when they see it. Putnam argued that the fact that we can have false thoughts about a category shows that we can think about something without knowing its defining features. Kripke’s and Putnam’s conclusion from these objections, i.e., their semantic externalism, has often been challenged, but most contemporary philosophers accept at least their arguments from error and ignorance (we can think about something even if we are wrong or ignorant about its properties). So, most philosophers agree that both arguments put at least severe constraints on definitionism and a theory of concepts in general.

Another serious problem for definitionism is the by now uncontroversial findings of so-called *typicality effects* in categorization and recognition tasks (Rosch, 1975). Rosch and her colleagues showed in a number of studies that subjects are faster at recognizing instances of a

category that possess more typical features of the category than instances that possess fewer typical features or properties. Furthermore, when asked to name members of a category, typical features are mentioned first and more reliably across different subjects. Strong versions of definitionism predict that subjects should be equally fast at recognizing typical and atypical members as long as they have the same necessary and jointly sufficient properties for category membership (Laurence & Margolis, 1999). However, this objection is only convincing if we assume that definitionism makes predictions about the speed at which participants make quick "rough and ready" categorizations.

Another interesting finding that became especially relevant in Jerry Fodor's work (e.g., 1998) is that words with supposedly more complex concepts as their meaning do not seem to be more difficult to process than simpler ones. If it is reasonable to assume that 'convince' is defined in terms of CAUSE TO BELIEVE, the word 'believe' should be easier to process than 'convince' since the one is constructed in terms of the other, meaning BELIEVE is the more primitive, less complex, concept. Such an effect however has not been found (Fodor and Garrett, 1975; Fodor et al., 1980; Fodor, 1998). The term 'believe' appears to be just as difficult to process as the term 'convince'. The absence of evidence is not evidence for the absence of such an effect. However, combined with the above findings and theoretical reasons, it is reasonable to assume that definitionism is widely rejected for good reasons – at least for a theory of how we apply our concepts and words. Still, the important role of definitions especially in science is not challenged here and the fact that complex concepts are definitional, in the sense of specifying descriptively to what kind of things they correctly apply to is not questioned in this thesis.

4.2 Prototype theory

To account for typicality effects and the fact that competent speakers do not seem to always represent necessary and sufficient application conditions for many of our ordinary language words, Rosh (1975) proposed that lexical concepts consist not of representations of necessary and sufficient properties associated with a category, but of a body of representations of typical properties ("typical features") associated with a category. According to this "prototype theory", the concept of dog, for example, would be identical to a body of features that are statically relevant but need not all be present in all instances of the category. Having four legs and being furry, for example, might be typical features of dogs and thus be part of the prototype of dogs. Still, she argues that it need not be the case that all dogs have to meet this condition. It would

be sufficient to only have a number of the represented features as long as this number reaches a certain threshold.

Note that according to Rosh (1999), typicality cannot simply be reduced to frequency. When asked whether a set of objects belongs to the category 'vegetable', not the objects that subjects were confronted with most frequently, such as onions or potatoes, were recognized faster. Instead, more seldom exemplars, such as cauliflower, carrots or peas displayed a processing advantage. Similarly, when asked whether a set of objects belongs to the category of clothes, pajamas and bathing suits were recognized more quickly as clothes than shoes, hats and gloves even though we wear shoes much more often than pajamas and bathing suits. This shows that even though frequency has an impact on categorization (more frequently retrieved concepts may still have a processing advantage), typicality is a better predictor of how easy we find it to reliably apply our concepts and words.

The main problem for prototype theory is that we often recognize instances of a category as a member of this category even if it does not have many or even any of its typical features. For example, Keil (1989) argued that prototype theory cannot account for the fact that we would not identify raccoons as skunks just by giving them a white stripe and a distinctive smell. This is at least difficult to explain by prototype theory. I call this a “causal-nomological effect” of categorization. It shows that we are able to recognize instances as belonging to a category based on more theory-like causal or nomological properties of the category. However, as noted by Edouard Machery (2009), prototype theorists are not committed to superficial, perceptually derived features. Still, it seems that even though prototypes can also include more structural, functional and abstract attributes, prototype theories are typically presented as involving mostly superficial, i.e., directly perceivable features (e.g., Margolis and Laurence, 1999; Prinz, 2002). In fact, superficial features do seem to give the best explanation for quick categorizations, and seem to be that which is needed to explain typicality effects.

Another supposedly major problem for prototype theory is that it cannot account for certain crucial properties of thought, especially its content and productivity. We might recognize dogs by means of their typical features but we would not say that the content of the concept of dog are the typical features of dogs or that DOG only applies to those entities that reach a certain threshold of typical features. This sounds similar to the point made by Keil, but it goes beyond it. It is not just that we identify objects as dogs even though they do not have any properties that are typical or frequent for a dog. The point is that when we think about dogs we think about

dogs in general and not just about those dogs with properties we take to be typical for them. It is thus not clear how our prototype of dog can be that in terms of which we use to think about dogs.

Moreover, as argued above, in order to explain how concepts compose, we need the meaning of the complex concept to be merely derived from the meaning of its part and some syntax rules. However, it seems that the typical properties of pets (furry mid-sized animals that live at home) and the typical properties of fish (large smooth animals living in the ocean) combined do not give us the typical properties of pet fish (see Fodor, 1998 for this example). There have been many attempts to avoid this problem (e.g., Hampton and Jönsson, 2012) but they all rely on discourse context and world knowledge to decide which features of a prototype are to be combined and which features ought to be added to get to the typical features of the complex concept, say of pet fish (see chapter 1 and 2, but also Machery and Lederer, 2012). So, on this picture what is needed to think about pet fish is that we already know what they are and which combination of features is the right one. It already requires us to have not just the concept of pet fish, but also knowledge about its typical features. What we need to explain concept combination is a more mechanical process that does not require us to already possess the concept we would like to get by combining simpler concepts. So, either concepts do not compose in the strong sense presupposed above, or prototypes are not concepts.

4.3 Theory Theory

Theory theorists aim to capture the finding of what I called "causal-nomological effects", i.e., that people know that some inner or more theoretical properties of an object or animal may be more decisive as to whether this object or animal belongs to a certain category than its typical features.[2] Theory theorists like Keil (1989) or Carey (2009) proposed that concepts are sets of *mini-theories* that store knowledge that is less superficial and more akin to scientific or folk-theories. Other theory theorists claim that concepts are only elements of theories, while yet others understand concepts as full-blooded theories (see, Machery 2009, p. 101).

One of the main objections to theory theory is that it is left underspecified. First, it is not clear what theory theorists mean by 'theory' due to the fact that there are many different notions of theory to choose from (Prinz, 2002). Theories in folk-psychology, for example, are very

[2] Note that theory theorists of categorization should not be confused with theory theorists in the debate on mindreading. However, there is a resemblance if we assume for example that people tend to rely on theory-like representations to decide whether a certain mental state applies to a person or not.

different compared to theories in neuroscience. While some theories are based on mechanistic models, others are based on simple causal laws or everyday observations of regularities. However, I take this objection to be rather tame. A theory can be considered merely a belief or set of beliefs about less superficial properties of a category and need not have much to do with anything like a proper theory in biology or chemistry. The core difference between theory theory and prototype theory is that only the latter stores statistically derived knowledge, while theories are supposed to store more essentialist, nomological, generic and functional information we learn, e.g., at school. Theories are thus not so much statistical, but resemble causal laws, such as, 'if a vehicle burns it might explode' or functional generalizations, such as, 'vehicles are used to move fast'. In addition, theories could be beliefs like 'the intention of the maker is important for the concept of art' or 'some animals have an unknown essence that make them the species they are' (Machery 2009, p. 103).

A more serious problem for theory theory is that theories alone might be insufficient to explain our everyday behavior. For example, a classic example of a theory is that water consist of H2O. This would qualify as a theoretical information in the sense required by theory theory. However, it does not allow us to pick out water in an everyday setting (we do not always carry a chemistry lab with us). Only in combination with another theory, say that H2O is responsible for its lack of taste and color, can we explain our more immediate reaction towards water. Thus, we could identify concepts with sets of theories. Such a set could be retrieved every time we interact with water and if asked about the nature of water it would be this knowledge that allowed us to give an educated answer. However, if theories contain knowledge that helps us identify typical instances of a theory it is not clear how it differs from prototype theory, which also posits a set of typical beliefs about an entity. So, theory theory may be best understood not as a competitor of prototype theory, but as an addition to it (as argued e.g., by Machery, 2009).

An additional problem for theory theory is that it seems difficult to find theories for all categories. What could be a theory representing PLAUSIBILITY or INFORMATION? Perhaps a theory representing 'plausibility' is “everything that is in accordance with our beliefs is plausible”. But this seems more like a definition than a theory. Moreover, this theory is circular because how do we identify assertions that are in accordance to our beliefs? Probably because they *appear* plausible to us. Perhaps, there are theories that can represent very abstract lexical categories, but, unfortunately, we hardly know them or find any of them uncontroversial, which is why philosophers still argue about them (e.g., in the case of ‘justice’ or ‘democracy’). Similarly, we could have a theory of the concept of love that can be expressed by the sentence

“people in love are usually nice to each other”. This might help us to identify people who are in love in some situations but not others and it is not clear whether this qualifies as a theory in the sense required by theory theory or rather as a belief about typical properties of a love-based relationship.

Finally, theory theory has the same problems as prototype theory when it comes to questions of content, strong compositionality and systematicity. It is not clear for example how theories represent the actual content of our concepts. Again, DOG is about dogs and not about a theory associated with dogs. Secondly, combining the theory of DOG with a theory of HUNGRY may not give us a theory representing the content of the complex concept of hungry dog (Laurence and Margolis, 1999). Again, it could be argued that we select some aspects of our theories about dogs and some aspects of our theories about things being hungry. However, again, such a selective combination probably requires us to consult discourse context and world knowledge and cannot merely be computed in a strongly compositional way as required by a theory of concepts.

4.4 Exemplar theory

The main difference between exemplar theory and theory theory, definitionism and prototype theory is that exemplars in the sense of exemplar theory are explicitly not supposed to be summary representations, i.e., representations that are supposed to “summarize” what all instances of a category have in common (Smith and Medin, 1981). Summary representations have, as we have seen, the major limitation that we usually find exceptions. As argued above, we can even find exceptions to correctly applying the word bachelor to unmarried adult males. Even more problematic are prototypes, which, almost by definition, exclude atypical exemplars simply if they do not possess the sufficient number of typical features.

Prototype theories have another major problem that has not yet been mentioned and that exemplar theory is a response to. Recall that one explanation of the way people represent categories is by representing a prototype which stores the typical features of that category. This would mean that a concept like FRUIT may contain both the features 'banana-shape' and 'red' because both are equally good predictors of something being a fruit. The worry is the following: prototype theory predicts that if we see a red banana we would as readily recognize it as a fruit as we would recognize an actual banana (for a similar example see Prinz 2002, 64). This is counter-intuitive because most people who see a red banana, due to its unusual color, would probably doubt whether it is a banana or fruit at all.

The counter-intuitive results of summary representations have led some psychologists to propose that concepts are not sets of typical features that are supposed to capture all or most instances of a category, but sets of actually perceived exemplars, i.e., representations of instances of the respective category (Medin & Schaffer, 1978; Brooks, 1978). In other words, instead of storing a set of features (i.e., representations of properties) that all or most fruits have in common, an exemplar stores representations about particular members of the category. Hence, it is predicted that whenever we encounter a new object, we compare this object to the various exemplars of fruits and if there is enough similarity between at least one of the exemplars, we conclude that this object is a member of the category.

Exemplar theories are often used in combination with prototype theories. For example, according to Prinz & Clark's (2004) hybrid model, an exemplar allows us to identify an unusual or atypical member of a category, while a prototype explains how we identify typical members. Another option is that one exemplar could constitute one individual concept (Machery, 2009), which means that one category would be represented by many different concepts.

However, as a single unified theory of concepts, exemplar theory has the same problems as its predecessors (Laurence and Margolis, 1999). First, it does not seem that the content of our concept of fruit is a set of exemplars of fruits. Instead, it seems that the content is the set of all fruits or the property of fruits. Secondly, it is not clear how exemplars combine without the help of discourse context and world knowledge (see again Machery and Lederer, 2012). Finally, it is not clear whether competent users associate exemplars with all concepts. Again, what is a good set of exemplars of the concept of plausibility or the concept of anti-radiation device? This does not mean that we cannot come up with exemplars for these categories, but that speakers may competently use words and concepts without representing exemplars.

4.5 Simulation theory

Simulation theory essentially claims that what underlies higher cognitive competences like categorization or linguistic understanding are simulations or models of, for example, concrete situations, individuals, objects and feelings that are based on representations of previous experiences or events. In other words, what eventually explains how I recognize something as a dog is a simulation or model of a dog in working memory based on certain beliefs of what I take dogs to look like that are stored in my long-term memory. This model of a dog is simulated in an *ad hoc* way and can be imagistic or realized by means of modal or amodal symbols (it is

not a theory about format). Once a simulation is generated, it is compared to a representation of what is being perceived, via some comparison operation.

According to simulation theory, what explains our higher cognitive abilities are two fundamentally different kinds, namely, what Barsalou (1999) calls "simulators" and "simulations". The former corresponds to a set of privileged information, i.e., information that has a processing advantage because it is frequently retrieved to apply a certain concept, associated with a category in long term memory. The latter denotes temporary simulations or models in working memory. The information stored in a simulator can be accessed partially to generate relevant simulations or models if context demands it. What goes into a simulator, i.e., which information per category is privileged is determined by the simulations that are regularly generated in certain contexts. What goes into a simulation depends on what is immediately available from the simulator, which in turn depends on standard psychological determiners of processing speed, such as typicality, frequency and context.

As with all theories of categorization in psychology, simulation theory's success will be measured by how well it can account for typicality, frequency, exemplar and causal-nomological effects. To explain especially typicality and frequency effects, the idea put forth by simulation theory is the following: to allow for a fast and accurate simulation, we would expect that, depending on the context or task at hand, some pieces of information in long term memory ought to be more readily available than other pieces of information. If all pieces of information were equally readily available in all contexts it would make the selection of the relevant information extremely difficult and slow. It is thus expected that the system structures long-term memory in a way that privileges information that is typically or frequently employed to generate or inform a certain situation model. Furthermore, we would expect a fast and reliable system to take context into account, besides frequency and typicality. Thus, representations of food should be more readily available in a restaurant context than in a school context simply because food has been simulated more in a restaurant context than in a school context. This means that simulations are not just generated based on information in long term memory, but that the organization of information in long term memory is fundamentally based on what has previously been simulated.

Simulation theory has thus at least the potential to explain regularity (typicality, frequency and exemplar) effects without positing prototypes or exemplars simply by acknowledging that some information is privileged based on previous simulations, typicality, frequency and context.

Causal-nomological effects that were especially well explained by theory theory can be accounted for either by including theoretical beliefs in the simulator or by referring to the flexibility of the simulators that can be supplemented with more theory-like world knowledge outside the simulator.

Simulation theory has a lot of empirical support. It is not possible to review all or even most of the evidence in favor of simulation theory here (see Bergen, 2012 and Feldman, 2008 for excellent summaries), so I will focus on findings that I find especially impressive and relevant for this thesis (especially chapter 3 and the conclusion). Rolf Zwaan and colleagues for example found that when participants are told a story, say about a flying bird, it takes them longer to determine whether the pictured item was mentioned in the story if the bird is depicted with spread as opposed to closed wings (even though the state of the wings were not mentioned in the story). Similarly, after hearing "John hammered the nail on the wall" processing slows down significantly when participants see a picture that shows a vertically oriented nail (Zwaan et. al., 2002). Processing slows down when participants see a nail in a horizontal position after hearing that John hammered the nail on the floor even though neither of the sentences made the orientation of the nail explicit.

Similar spatial effects have been found for action verbs like 'kick' (Pulvermüller, 2013), abstract nouns like 'slave' and 'master' (Schubert, 2005) and even for mathematical concepts, which have long been considered paradigmatic examples of concepts that are detached from perception (for an overview see Fias & Fischer, 2005). For example, a number of studies have been able to demonstrate associations between spatial and numerical representations. To name just one, Dehaene et. al. (1993) have shown that subjects respond to smaller numbers faster when they are presented spatially to their left, while larger numbers are processed faster when presented to their right (the so-called the *SNARC effect*).

Even in the case of negation, one of the key examples frequently raised against simulation theories of linguistic understanding (see again the commentary to Barsalou, 1999 or Dove 2009), it has been argued that simulations are at least relevant. For instance, in a semantic priming study, Giora et al. (2004) found that associated concepts of the negated concept were just as readily available as associated concepts of the affirmed concept (e.g., "The instrument was sharp "versus "The instrument was not sharp"). Similarly, a series of studies, Kaup et al. (2006, 2007) found that affirming the sentence "The umbrella is not open", when seeing a picture of a closed umbrella, is facilitated after 1500ms, but not after 650 ms. Both sets of

studies suggest that not only do we engage in the simulation of the positive scenario of the negated sentence, but that we also represent the negated scenario suggesting a two-step process of first representing the positive and then the negative scenario (but see Tian et al., 2010).

The idea that we simulate models based on world knowledge and not on what is made explicit in a natural language sentence is also suggested by ERP evidence reporting a brain response for unexpected stimuli at around 400ms after word onset, called the "N400". One especially impressive finding is that such a N400 can be found in Dutch speakers who hear the sentence "trains are white" considering that trains are usually yellow in the Netherlands (Hagoort et al., 2004). There is nothing in the meaning of train that would make it probable that trains are yellow, but it seems that merely mentioning the term 'trains' to Dutch speaker generates expectations of yellow trains that is violated (causing an N400) if the sentence specifies a different color.

That we simulate concrete situations using very concrete representations even for very abstract terms, has found some support in the recent literature on so-called "situated cognition". One of the first studies in this line of research by Wiemer-Hastings & Xu's (2005) reports evidence that subjects retrieve situational, social and emotional information for abstract categories like 'emancipation' or 'freedom'. This suggests that abstract concepts are at least causally linked to emotions and social situations.

Finally, another psycholinguistic variable called "context availability" seems to play an important role in early unconscious word processing, especially of abstract words. It has been suggested that words are more difficult to recognize and understand if it is difficult to generate contexts or situations in which this word could be used (Schwanenflugel, 1992). This applies especially to unfamiliar and abstract words like 'decency' or 'allow' (based on Altarriba et al, 1999). However, once these words are presented in context, the processing delay disappears (Schwanenflugel, 1988; Kousta et al., 2011). This suggests that the ability to simulate situations based on basic associations triggered by an expression strongly constraints our language use.

So, the simulation theory has a lot of empirical support and theoretical advantages. The main challenge to simulation theory are abstract concepts (Dove, 2009; 2011). The idea is that we cannot represent abstract concepts like TRUTH or KNOWLEDGE in terms of concrete situations because abstract concepts, by definition, do not refer to anything physical that could be simulated in terms of a representation of a concrete particular. I argue against this objection in chapter 3 and the conclusion. A second problem is that it is not clear what exactly the concept

is supposed to be, i.e., whether it should be the simulator or the simulation? A third set of problems are stability and again its difficulty to account for compositionality and systematicity without the need of discourse context and a more fundamental level of representation that ambiguity could be conceptualized and processed in terms of. It does seem that by its very nature simulations and also simulators are highly context dependent and at least unstable across even a short period of time.

Finally, simulations do not seem to represent the content of our concepts. The concept of dog is about dogs and neither about certain simulations of dogs nor about what these simulations resemble. Furthermore, often, we possess a concept without being able to generate any simulations or at least no simulations that have anything to do with what the concept is about. This goes back to the problem mentioned above raised by Putnam that we may be able to think about something while making mistakes about what it is really like. Moreover, it is not clear how simulations can compose in an unambiguous and context independent way as required for a theory of concepts in the sense introduced above. For example, it is not clear how a simulation of a pet can conjoin with a simulation of a fish to generate the concept of a pet fish unless we allow context and background information to adjust the resulting simulation.

4.6 Concept Atomism

All theories so far assumed that most lexical concepts are complex entities that either consist of a set of necessary and sufficient or typical or causal features, or sets of representations of exemplars. However, all of these theories also have to assume that some concepts turn out to be simple, i.e., not further analyzable, even if these may not be lexicalized. If we never reached a "rock-bottom" layer of representation, we would not be able to really specify the meaning of a given concept. Take the concept associated with the word 'dog'. This concept may be understood in terms of features representing *furry* and *bark*. If our concept of furry is itself understood in terms of other concepts and these again in terms of other concepts we would not reach an understanding of the concept of dog or any concept without ending up in a vicious circle whereby the existence of one concept depends on the existence of all the another concepts (see Fodor, 1998 and Fodor & Lepore, 1992).

Concept atomism is the empirical hypothesis that most lexicalized concepts (i.e., concepts that received a lexical expression) are atomistic, i.e., not complex (not composed of simpler concepts). This does not mean that concept atomists claim that all lexicalized concepts have to be atoms. For example, if it turned out that 'bachelor' is best analyzed in terms of the complex

concept UNMARRIED MALE, then the concept of bachelor is of course not an atom, according to atomists. Fodor, the most prominent concept atomist, argued that the question of which lexical concepts are atoms and which are complex is an empirical one and he put forth empirical evidence to support concept atomism (e.g., Fodor, 1998). Furthermore, concept atomists also do not argue that most concepts are atoms. In particular, all complex concepts are not atoms, by definition. Since arguably most concepts are complex concepts (Fodor put much effort into showing how concepts could be combined), concept atomists usually think that most concepts are *not* atoms (but, again, they think that most *lexicalized* concepts *are* atoms).

Atomism is rejected by most psychologists for one obvious reason: if a lot more concepts are simple or atomistic than we thought, especially the ones used as building blocks of language, then how can they possibly explain how we make categorizations or exchange thoughts with others? Categorization cannot work in most cases unless we allow for at least some set of beliefs or representations by means of which we can judge whether an object belongs to this category. For example, assuming we identify tomatoes by means of their superficial features, how can concept atomists explain how we can use our concept of tomato to identify objects as tomatoes (see Prinz, 2002, 99 for this objection)? If atomism cannot account for categorization and inference-making it cannot be a theory of concepts in the sense important for psychology.

Similarly, if we take seriously the idea that concepts ought to explain other conceptual abilities, such as decision making or inference making, we need concepts to guide our decisions by storing certain beliefs about the world. To make such decisions as whether we would prefer living in a democracy or an autocracy, we need our concepts to specify what exactly it is for a country to be a democracy or an autocracy. Since atomism does not allow that lexicalized concepts are constituted by features representing properties that help us understand what distinguishes autocracies from democracies, it is difficult to see how it could explain our decision-making behavior regarding, say, preferred forms of government in terms of our concepts of democracy or autocracy (it may however turn to a different entity for this explanandum).

Another reason why atomism is rejected by most psychologists and also many philosophers today is that its explanation of how we acquire concepts is widely taken to be implausible (Prinz, 2002). According to atomism, even such concepts as ART or DEMOCRACY cannot be learned, at least by means of learning a number of more fundamental concepts that we can then learn to combine to form the new complex concept. If even concepts like ART or DEMOCRACY are

supposed to be atoms then how can we learn them? To see why atomism cannot explain concept acquisition well, at least according to its opponents, consider how feature-based theories explain the acquisition of concepts. According to these views, concepts are constituted of features, which only at the lowest level will be primitive. These primitive concepts may be derived from the physical interaction with our environment or can be innate. Color-terms or shapes for example may be derived from interacting with different shapes and colors that we may detect via innate shape or color-detectors. If concepts are those mental representations that explain how we detect their referents, these detectors can be identified with the respective concept (Prinz, 2002). However, if most of our lexicalized, i.e., most of our most important and frequently used concepts are primitive, how do we acquire the ones that cannot easily be derived from the environment, as in the case of ART or DEMOCRACY?

On the other hand, atomism has a lot of advantages that speak especially to theorists of concepts with more philosophical leanings. Atomism can explain why so few concepts can be sufficiently analyzed, i.e., why conceptual analysis in philosophy in the sense of finding necessary and sufficient conditions has largely failed to be successful. Secondly, it allows for a rather simple account of compositionality and systematicity. If PET means *pet* and FISH simply means *fish* then pet fish simply refers to pet fish in a strongly compositional way. Both compositionality and systematicity then accounts for our sense of productivity of thought according to which we can produce an unlimited number of new thoughts by combining concepts in systematic ways.

4.7 Pluralism

While atomism and definitionism cannot account for categorization and how we make inferences, they work rather well as theories of conceptual representations. Both (atomism for lexicalized and definitionism for complex concepts) explain the most important properties of thought, especially compositionality and systematicity which are important to explain productivity. Prototype theory, exemplar theory, theory theory and simulation theory, on the other hand, are especially well equipped to account for categorization and inference making, but fail to account for the properties of thought from providing a theory of content to productivity.

Many contemporary philosophers of language and psychology have come to realize that there is potential for combining atomism and definitionism with prototype theory, exemplar theory, theory theory or simulation theory. In fact, many have argued that concepts might have two kinds of content, an atomistic or definitional content that explains compositionality,

systematicity and productivity, and another aspect or kind of content explaining categorization and inference-making. Such "dual-content views" have been proposed for example by Prinz, (2002), Weiskopf (2009), Del Pinal (2015) or Vicente (2018), but they can also be attributed to earlier views by Frege (1948) as well as theories using the mental files metaphor most thoroughly explored by Recanati (2012).

Note that there are different kinds of pluralist views that must not be conflated. Most psychologists today are pluralists in the sense that they believe that we may have different kinds of representations to explain categorization. They accept that prototype theory, for example, may not be sufficient to explain all explananda with respect to categorization and that we may need to also assume that we have additionally exemplars and theories, perhaps even mental simulations, to explain all of human and non-human higher cognitive capabilities (see Machery, 2009 or Weiskopf, 2009). Another kind of pluralism combines psychologically successful theories like prototype and exemplar theory with concept atomism or definitionism. Only if this latter view also includes the assumption that the same concept can have two contents, is it called a "dual-content view". According to this kind of pluralism, a concept consists of two kinds of content: an atomistic or definitional core that accounts for concept stability, content and compositionality and a second content explaining how we apply this concept consisting of prototypes, exemplars and theories that are associated with this core as part of a single concept.

A much less explored and currently much less popular kind of pluralism and way to combine the virtues of psychological theories with definitionism and concept atomism is to simply say that concepts are not the kind of things that need to explain categorization. Instead, concepts are merely the kind of things we think in terms of. Such kinds might not need to explain how we in fact apply concepts to objects. This approach would have it that we make a strict distinction between concepts and the structures we use to apply concepts to objects in categorization or decision-making, which, again, we may simply call "categorization devices" or "abduction devices". Much of this thesis argues for this latter far less explored (but arguably more traditional, Putnam, 1975; Rey, 1983) pluralism as a way to make sense of the advantages and disadvantages of the theories of concepts summarized above.

In fact, it is not clear why we would want concepts to explain *how* we a) correctly and b) actually apply themselves. This, today at least, is typically taken for granted as a desideratum rather than fully defended (e.g., Prinz, 2002). *Prima facie* the idea that concepts ought to explain categorization is of course natural. We can easily make sense of this idea if we consider that

concepts are supposed to be just a special kind of representation with certain properties that explain thought. For example, we would assume the same double role for other kinds of representation, in particular pictures. We would assume for example that we decide what a painting is a painting of, i.e., what it correctly applies to is directly indicated by its structural features, i.e., the represented shapes and colors. We usually decide whether the painting of Mona Lisa picks out Mona Lisa by looking for similarities between the person and the object depicted. Many people would at least be surprised to hear that a picture that contains two dots and a line is supposed to be a picture of Donald Trump.

Compared to other kinds of representation, however, the idea that concepts ought to explain categorization comes much less natural. For example, we would not assume that words are not only that which we apply to objects in sentences, but that which we apply words in terms of. It is simply absurd to think that we use the word 'table' to apply itself, i.e., to apply the word 'table' to tables. We do use the word 'table' to denote tables, but we do not use the word to decide whether we should apply it to tables or not. We use our knowledge of what tables are usually like and other information about tables to make this decision. So, if it is hard to make sense of what it could even mean for words to be not only those things we apply to objects but also those things that tell us *when* we should apply it to objects, why think that we use the concept of table to decide whether something is a table, i.e., to apply the concept of table to tables? The question may therefore be whether concepts are more like realistic pictures that represent via depicting similarities or more like abstract words that represent, e.g., by convention?

The idea that my concept of a table ought to explain not just how I think about tables but also how I identify something as a table also comes from a tradition in the philosophy of concepts according to which we can analyze a concept by means of what it takes to possess this concept, i.e., by means of studying its possession conditions (e.g., Peacocke, 1989). It is assumed in this literature that concept possession is best understood in terms of abilities, e.g., to correctly classify something as falling under a specific concept or as finding a given concept application as immediately intuitively correct. This approach is of course not absurd. In fact, it is difficult, but not impossible, to make sense of the idea that a person may possess the concept of water without at all being able to identify things as water or have any beliefs about water whatsoever. However, many engaged in the literature on meaning and concepts believe that exactly this is the message we should take from Putnam's and Kripke's arguments from error and ignorance.

To understand the argument put forth in this thesis it is important to emphasize that I *do not* take a stance on the debate on whether we ought to find certain inferences or concept application as immediately adequate in order to possess a concept. This question has been thoroughly discussed in the past especially by Fodor (1998, 2004) and shall not be discussed in detail here. I only commit to the idea that a radical holism according to which all concepts are individuated by means of other concepts is probably false considering that this would make it difficult to see how we can really account for the stability of concepts (which I take to be the most important property of concepts) and in particular successful communication. So, all I commit myself to here is that at least some concepts are conceptual atoms, say concepts of shapes or colors. My argument goes beyond this traditional debate. The attempt is here to show that, whichever view about concept possession and individuation we choose, the question of how we in fact apply our concepts (which beliefs we in fact use to decide whether a given concept applies), still needs to be settled. As we will see, this approach relies on a much weaker version of the argument from error and ignorance that is compatible with thinking about concepts as being individuated by means of inferences we find immediately intuitive.

Another reason why it is widely assumed that we require concepts to store more than an atomistic core that especially convinced Prinz (2002) to put forth a dual content view is the problem of Frege cases, i.e., the problem that we can have contradictory beliefs about the same reference without knowing it. Prinz assumes that we have to include beliefs in our concepts to explain why Louis Lane can rationally believe conflicting things about Clark Kent (he is weak and boring) and Superman (he is strong and exciting) without knowing that she has different beliefs about the same man, even after reflecting about this possibility. I will not discuss this point here further, but I will come back to it in the conclusion where I propose an alternative way to make sense of Frege cases without assuming dual contents by employing my notion of a categorization device.

Another intriguing hypothesis why most current philosophers of concepts assume that concepts have to explain categorization is a simple misunderstanding of Fodor's (1998) desiderata for a successful theory of concepts. It is difficult to exaggerate the impact of this book. *Concepts* has been cited more than 2500 times so far, which is more than any other contemporary philosophical book on concepts that I am aware of (Peacocke's *A Study of concepts* has been cited just about 2000 times). Even psychologists interested in concepts have taken his criticism about prototype theory expressed in this book seriously, which resulted in a number of attempts to explain how prototype theory can explain compositionality (e.g., Hampton & Jönsson, 2012).

In Fodor (1998, p. 24), we find an influential list of desiderata for a theory of concepts that heavily informed Prinz's (2002) list of desiderata. This list contains the same desiderata that I have introduced above. In addition, it states the following:

> Concepts are categories and are routinely employed as such. To say that concepts are categories is to say that they apply to things in the world; things in the world 'fall under them'. So, for example, Greycat the cat, but not Dumbo the elephant, falls under the concept CAT. Which, for present purposes, is equivalent to saying that Greycat is in the extension of CAT, that 'Greycat is a cat' is true, and that 'is a cat' is true of Greycat.

This quote can easily be misunderstood as saying that concepts have to explain how we recognize cats as cats or Dumbo as an elephant. However, this is not what Fodor is arguing here. Fodor is not arguing that concepts need to explain categorization. He simply states that concepts need to be the kind of things we can apply to objects. They are not the things, Fodor later argues in his book, that explain how we can make correct categorizations, they are the things that we apply *in categorization*. To apply a concept just means to make a categorization. Fodor argues that we use beliefs to apply concepts and that this is identical to categorization. Fodor does *not* argue that the concept of x is used to apply the concept of x.

5 Why abstract concepts?

This thesis is not *just* about concepts in general. It is also a thesis about abstract concepts in particular. Abstract concepts are, very roughly speaking, concepts that refer to things that we cannot directly see or touch. They are not to be confused with abstract objects or even concepts of abstract objects. A more technical definition has been proven difficult to come by in the philosophical and psycholinguistics literature and is not needed for the purposes of this thesis (it is part of future work). The best way to get a feeling for the distinction between abstract and concrete concepts, for now, is by means of examples. Paradigmatic abstract concepts are the concepts of democracy, love or happiness. Less paradigmatic and less clear cases are the concepts of vegetable, electron, mother, red or president. Paradigmatic concrete concepts are the concept of table, tiger and water.

Abstract concepts are fascinating for many reasons some of which shall be summarized in this section. However, despite their fascinating nature, abstract concepts are hopelessly understudied in philosophy, linguistics and psychology. While there are many books discussing

concepts with a focus on concrete concepts, as well as many books discussing specific abstract concepts, like the concept of truth or the concept of knowledge, there is, to my knowledge, no philosophical monograph that specifically deals with abstract concepts as such or that gives a theory of what abstract concepts are or how they are acquired.

The first thing to appreciate when it comes to abstract concepts and words is just how important they are to us. They not just incredibly important especially to a philosopher who spends most of their professional life using, discussing, describing and analyzing abstract concepts, but to humanity in general. Human adults and even children use an abundance of abstract concepts almost all the time and a lot of speech acts that make our social lives worth living are abstract, from pledging eternal friendship to baptizing a child. It is thus very curious why especially philosophers have not published more about the nature of abstract concepts and the way they are used by our minds considering that without them the very practice of doing philosophy would be not just impossible but unthinkable.

Note also that abstract concepts are usually much more important to us than concrete concepts. The focus on concrete concepts both in philosophy of mind and psychology is thus even surprising considering that normally we do not really care about them. We care about abstract concepts. We do not just care about what these concepts pick out, but even non-philosophers care about the concepts themselves. We do not just care about love, racism, democracy, freedom or equality, we care about the concepts of them. This contrasts them with concrete concepts. We do not care about the concepts of water, gold or cat even though we care about water, gold and cats. What we mean by the term 'gold', 'water' or 'cat' will probably not suffice for a good evening TV program or Youtube channel. The concepts of justice, racism or inequality do. So, why have philosophers of language and psychology focused on giving us a theory of concrete concepts if, especially as philosophers, we do not really care about them?

The second thing to note is that abstract concepts are not just essential for us to think, but crucial for the way we experience life. For example, the French writer François de La Rochefoucauld once said in his *Reflections* (maxim 136) that people would never fall in love if they had not heard of the existence of such a thing. I think that this is probably true. As a highly speculative but nonetheless intriguing illustration of this idea consider Aldous Huxley's dystopian novel *Brave New World* where people have difficulties even grasping the concept of love. The interesting fact about this world is not that the people in it have been explicitly discouraged from engaging in meaningful romantic relationships. One point of the book is that most of them

do not see any reason in doing so even without the discouragement. They did not see a reason to acquire or apply the concept in the first place. While the creators of the world did have the concept of love, they set up the world in such a way that ordinary citizens would not come into the position of imagining or even wanting such a thing as love. What this suggests is that if it was not for the arguably culture specific concept or concepts of love we would interpret certain biological fundamental feelings, such as the attraction to another person, very differently, perhaps in the way described in brave new world. If it was not for abstract concepts, we would experience life very differently and live our lives in fundamentally different ways.

Third, the analysis of the psychology and nature of abstract concepts has an immediate real-world application that has especially in recent years increased dramatically in importance. This application and one of my main motivations for writing this thesis is identity politics, i.e., the publicly held discussion on who holds the authority over a person's social identity (see chapter 4). Especially in the United States of America this debate is held very actively, fiercely resulting very often in physical and psychological violence (shootings being only the tip of the iceberg). Identity politics affects the lives of us all in often unnoticed but extremely deep and fundamental ways. It concerns the question of who has the authority over who we essentially are. As a particularly interesting example consider the question of whether homosexual couples can marry. This debate is usually not about the nature of marriage, but, first of all, about the very concept of marriage. One of the strongest arguments that opponents of gay marriage have is that the concept of marriage simply does not correctly apply to homosexual couples. So, many same-sex marriage opponents argue that they would allow homosexuals to “marry” but only if this marriage was recognized as something different, i.e., if there is a clear conceptual distinction between same-sex and different-sex marriages.

Another example that I find even more fundamental to many people's experience are questions of the correct application of concepts pertaining to gender, i.e., concepts of woman, man, non-binary or gender fluid. More and more people today come forward claiming that their assigned gender is different from the gender they identify as. Transgender issues affect a significant number of people and lack of acceptance in our society causes real trauma and suffering in this especially vulnerable fraction of society. Sadly, the rate of suicides among transgender people ranks among the highest, making trans issues a real and pressing health problem (not to mention the frequent violent attacks against the trans community). There remains much hostility in all societies on earth towards trans people and much of this hostility is based on conceptual issues, namely that in our societies we apply the concept of man and woman in highly restricted and

essentialist ways. (Note that other concepts that deserve just as much attention are concepts of race, age, class etc.).

One might object that gender issues are not conceptual but metaphysical issues. However, it cannot be stressed enough that before we can investigate the nature of anything we have to be able to think about it, at least in some rough way. For example, before we can study the nature of woman we have to make some conceptual commitments as to whether we are investigating a biological kind, or a social kind. The answer we give to such a conceptual question affects also who is responsible for studying these issues. For example, if we believe that our concept of woman is more like the concept of water, then we ought to do empirical studies in biology to figure out what it takes to be a woman. If we think that the concept of woman is more like the concept of president or money, then we should give sociologists the authority. If we think that the concept of woman is more like the concept of democracy or art, then it is largely on us to decide on the right description to prescribe what it takes to be a woman. So, to sufficiently analyze conservative arguments that claim to have an authority over what counts as a marriage or who is a woman or a man, we first require an account not just of the concepts of marriage, woman, man and so forth, we need at least some basic understanding of the nature of concepts and abstract concepts in general.

Another objection that partly takes us back to the idea that important concepts can be understood in terms of a family resemblance relation is a kind of conceptual relativism according to which everyone might associate different concepts with their linguistic expressions and that there is no matter of fact about which concept is the right one and that as long as we can successfully communicate it does not matter. However, this lax attitude is problematic. We want people who say that trans-women are not women, that homosexual couples cannot, by definition, get married to be either right or wrong about these issues. However, in order for there to be the possibility of rightness or wrongness, we have to agree on the using the same concepts.

A fourth major motivation to study abstract concepts is that there has recently been a rather large body of research on abstract concepts in psycholinguistics that one can critically assess in light of methodology and content, and in light of what these results tell us about abstract thought and language in general. In other words, a thesis on abstract concepts, especially one that discusses the relevant empirical work, would have been much less interesting 40 or even 20

years ago. This recent increase in scientific interest is partly due to the recent increase in research on embodied/situated/4E cognition that will be discussed below (chapter 3).

By 'situated cognition' or '4E cognition' I refer to a general program in cognitive science to look for cognition in, for the tradition, unusual places, i.e., actions, the body, the context or situations, the past, the future, the motor cortex, the visual cortex, culture and so on, i.e., anything that does not, very roughly speaking, reside in the pre-frontal or temporal cortex, or that is not reducible to traditional amodal computational processes. Considering that more traditional approaches would not predict a large difference between abstract and concrete concepts in terms of mode of representation (for the tradition all concepts are essentially non-perceptual) they would not assume that abstract concepts need any special attention. Thus, although abstract concepts are often considered a problem or even a refutation of certain strong views of situated cognition, it has also, ironically perhaps, motivated much of the recent interest in abstract concepts that does, intuitively speaking, predict a difference.

This brings me to a final key motivation to study abstract concepts, namely the idea that they pose challenges to currently available theories of concepts. A theory of concepts that cannot account for abstract concepts cannot be a full-blown theory of concepts. Any complete theory of concepts must have a wide scope. Since more traditional (but still contemporary) theories of concepts, proposed for example by Fodor (1998) or Peacocke (1992), have paid very little attention to abstract concepts it is fruitful to investigate whether these theories of concepts are even applicable to abstract concepts. Similarly, a more recent, but particularly influential, theory of concepts in psychology is that concepts are simulations or simulators (Barsalou, 1999) of concrete perceptual experiences grounded in perceptual representations. Abstract concepts, i.e., concepts that do not refer to a physical object we can directly perceive, are of course a problem for such a theory and the strongest objections to such a view are based on abstract concepts (see chapter 3).

6 The method and claim of this thesis

I think the best strategy for the early days of theorizing about the nature and psychology of abstract concepts is to simply try to apply the theories designed to explain concrete concepts to abstract concepts and to see how far we get. This is the strategy of this thesis. The idea is to try to make sense of the theories of concepts we have (chapters 1 and 2) and then apply them to abstract concepts (chapters 3 and 4). The hypothesis is that abstract concepts are not really so different from concrete concepts. At least I argue that they are much less different than one

would pre-theoretically think and that our intuition that they are different is based on misguided assumptions about semantics and its relation to psychology. If abstract concepts are not importantly different from concrete concepts, then it should be easy to explain them by means of the same tools we have to explain concrete concepts. In other words, if a situated or empiricist account of concrete concepts sounds plausible then it should also be possible to theoretically account for abstract concepts in an empiricist or situated framework.

So, one hypothesis I would like to propose and defend here is that a good theory of concepts in general will tell us everything we need to explain abstract concepts. This of course means that in order to give a theory of abstract concepts, all we need is a good theory of concepts. The immediate problem for this approach is that we currently lack agreement in the theoretical or empirical literature on concepts and that even fully developed theories of concepts are sparse. However, in my view, there is far less *real* disagreement in the theoretical literature than is often assumed. To work out this agreement in detail is not necessary for the main purpose of this thesis, so the more modest strategy that I shall pursue here is to develop my theory of concepts around or independently of key controversies in the literature. I think the main philosophical controversy surrounds the meta-semantic question of what determines the content of a concept. As the reader will see, I will raise most of the points in a way that is independent of what stances one takes on foundational issues in the semantics of the mental.

In my view, the real disagreements on the nature and application of concepts lie primarily in the empirical literature. However, I here try to shy away from any empirical claims or predictions as much as I can (with some exception). When referring to empirical studies, one is usually advised to rely mostly on review papers as opposed to merely individual empirical papers. Unfortunately, often such papers are not available, so instead I will try to give a more detailed analysis of a number of key studies in psychology and neuroscience. However, a brief disclaimer is still necessary: the reader will notice that in the following I defend a (very roughly) empiricist view of concepts as proposed for example by Lawrence Barsalou (1999) or Jesse Prinz (2002). However, this defense is only a defense of *the plausibility* of such a theory of concepts and I do not rule out alternatives. The main aim is to do away with certain misunderstandings that stand in the way of the situated framework in order to allow this theory of concepts to account for abstract concepts. I do not here give or aim at a full account of such a theory.

The main method of this thesis is to identify conceptual confusions, e.g., where researchers fail to make an important distinction or where they use the same expression in different senses (i.e., using different concepts). I then try to disentangle these confusions and point out positive consequences of this disentanglement. In this thesis, I argue that the main confusion in the interdisciplinary literature on concepts, one that stands in the way of applying a situated theory of concept to abstract concepts, is a conflation of a concept with the structures we rely on when applying this concept. In addition, it is often assumed that that which we use to apply a given concept *correctly* also ought to tell us how the content of this concept is determined.

I argue that it is unlikely that our concepts are identical to that by means of which we apply our concepts. The only way this *might* be possible is if the content of our concepts would also be that which we use to apply this concept. For example, if the content of the concept associated with the word 'tomato' would just be the complex concept RED ROUND FRUIT, meaning that the concept of tomato just is RED ROUND FRUIT and we also only apply our concept of tomato to red round fruit. If this were the case, we would be able to explain how categorization works by means of semantic content. However, even in such a case we have to distinguish correct from actual concept application. We do not correctly apply our concept of tomato by means of its content, we apply RED ROUND FRUIT because of our background knowledge of diagnostic features of RED, ROUND and FRUIT.

None of this means that we do not rely on any conditions of applications to apply our concepts. Still, it is crucial to acknowledge that the conditions of *correct* application and the beliefs we *actually* use on a daily basis to decide whether a concept applies, need not be identical. In fact, a general answer for what it is that explains my use of an abstract word is readily and surprisingly easily found: what drives linguistic understanding and language use especially of abstract concepts are diagnostic beliefs stored in prototypes, exemplars, theories or simulations. These are representations that provide some conditions of applications, but may not determine the content of the respective concept and are not to be understood as necessary nor sufficient conditions for concept application. As argued above, the main idea behind all these theories is that we apply our concepts and words not just by means of necessary or sufficient conditions, but by means of all kinds of relevant world knowledge. For example, these theories predict that we apply our word 'dog' and our concept of dogs when we see something that looks like a dog because we know what dogs typically look like or that we apply our concept of table if we see something that roughly fulfills the function of a table.

I see no reason to *fundamentally* disagree with the answers provided in psychology (even though I try to make them more precise). Again, they all boil down to the same thing: concept and word application is usually based on diagnostic features and not just on necessary ones. However, so far it has not been fully understood in the concept literature what this exactly means and it has so far not been properly applied to questions in linguistics and even psycholinguistics. The reason why it has not been properly integrated in such theories is, I think, a very deep and largely underestimated one. The reason is that the idea that we apply concepts by means of diagnostic features and not by means of necessary conditions is difficult to bring in accordance with deep-rooted and largely uncontroversial requirements for a theory of thought, i.e., cognition and language, that have been introduced above. These requirements are not just philosopher's inventions. They are just as based on empirical observation as anything else in cognitive science. So, they cannot easily be dismissed.

Nobody has made these requirements for a successful theory of concepts more explicit than Jerry Fodor. The recently deceased American philosopher who is often introduced as one of the fathers of cognitive science pointed out that the problem with diagnostic features is that they fail as explanantia for key explananda for a theory of thought and language. Again, such a theory requires that conceptual representations can be compositional and systematic in a strong context-independent sense and need to be about the kind of things we intuitively think our mental states are about. Moreover, they need to be shared in a very strong sense, so that two people can have the identical concepts and not just similar ones. Nobody has yet been able to bring the entities that psychologists tell us we use to speak and think in line with the kind of representations philosophers of mind and language tell us we think *in terms of*. I think that the present thesis makes an important and crucial contribution towards this goal.

7 Overview of chapters, theory and results

This dissertation has two parts. In the first part, I lay out very general theory of concepts that can be supplemented in various ways depending on the more specific theoretical commitments of the reader. The aim is merely to say what concepts are in the most general and I hope largely uncontroversial way. In particular, I distinguish between two notions of concept, one most relevant for philosophy and one most relevant for psychology. I call the former 'concept' and the latter 'categorization device'. In the second part, I apply the general theory of concepts to a specific type of concepts, namely abstract concepts. The general hypothesis is that abstract

concepts are not special and that the intuition that they are is based on nothing more than vague intuitions that are easily debunked.

The general theory of concepts I propose is rather simple. There are two kinds of concept. Complex concepts like RED COFFEE and simple concepts (atomistic concept) like, arguably, WATER. The content of complex concepts is descriptively determined, i.e., by the description consisting of simpler concepts. Abstract concepts that are descriptive are easy to explain. All we need is a description to narrow down what exactly we aim to think about, as I have done with the notion of concept above (I gave a description of the functional role of concepts). To construe a concept of democracy for instance all we have to do combine simple concepts to generate a description that specifies necessary and sufficient features of a democracy. Since it is unlikely that all uses of the word 'democracy' can be explained by means of a single description or complex concept, we may generate a number of concepts of democracy. This however has no theoretical disadvantage and, in my view, accurately describes common practice in the natural and social sciences. Natural language is highly polysemous in this sense.

Then there are non-descriptive simple concepts, i.e., concepts whose content cannot be determined by means of a description. With atomism, I agree that many ordinary language terms are associated with concepts that are non-descriptive, i.e., they cannot be analyzed, i.e., reduced to a single set of simpler concepts. However, I disagree with atomism that most lexical forms or expressions are associated with a single primitive concept. Instead, I take a more pluralist stance according to which the use of many natural language expressions is best interpreted by means of a number of both primitive and complex concepts. Again, I take it that from all we know from psychology, linguistic behavior and conceptual behavior are mostly driven by psychological entities like prototypes or simulations and not by a single concept. The content of non-descriptive concepts has to be determined at least partly by the world. This means that the beliefs we have about the application conditions of such concepts may be incomplete or false.

As I explain in chapter 4, in order for content determination of conceptual atoms to work we need the contents to be mind-independent, stable, real kinds. Here, I also argue that more kinds meet this requirement than we might think. In particular many abstract social kinds are like this, especially *woman*, *white*, *heterosexual*. The question now is how we establish and sustain a relation to such kinds? This is a problem when we think about abstract concepts as referring to kinds we have no direct perceptual contact to, i.e., those things we cannot touch and directly

see. I argue that we can make sense of abstract cognition if we think of abstract kinds in terms of Boyd's (1999) homeostatic property cluster theory of scientific kinds. The idea is that we can relate to a real kind by detecting the superficial properties that it correlates with (Margolis, 1998). This is one part of my solution to the problem of abstract concepts for a theory of situated cognition. If it is successful, we would have a good theory of how we get to acquire and apply abstract concepts in a naturalistic and plausible world related fashion.

In chapter 1, I introduce the notion of concept in philosophy and psychology. I argue that the uses of the term are importantly different, i.e., that both use different notions of concept. This work is published in *Synthese*. In particular, I make a distinction between concepts and categorization devices. I argue that the former is that which we apply to objects and the latter denotes the sets of beliefs that we apply concepts with, i.e., that we use to decide whether a given concept applies or not. I argue that this distinction solves at least one problem that has loomed large over the interdisciplinary concept literature: the kinds of concepts that psychologists argue are used in categorization, what I call categorization devices, cannot compose in a strong sense in which concepts ought to compose, i.e., independent of context and speaker intentions. Again, the problem here was that concepts ought to be the "bottom layer of thought" (my expression), which means that they cannot be interpreted by means of a more fundamental level of representation. I argue that this problem is based on the content-categorization confusion and that categorization devices, if they are not concepts, need not compose in the strong sense as concepts need to compose.

In chapter 2, I argue that there is no conceptual or empirical reason to assume that the notion of categorization device, which in this chapter I call 'concept' (in accordance with the relevant literature), picks out a context-insensitive set of mental representations. This work was published in *Philosophical Psychology*. The main argument is that the empirical evidence that could possibly speak in favor of at least the most developed account of "invariantism" by Machery's (2009) can also speak in favor of "contextualism". It thus seems that both views eventually collapse into each other. However, contextualism has a methodological advantage. I argue that it better accounts for the compositionality of categorization devices and that it can better account for how we apply abstract concepts, in particular superordinate concepts like ANIMAL or VEGETABLE.

In chapter 3, I defend the simulation theory against its most challenging objection, namely that they cannot account for abstract concepts. Here I apply the concept-categorization device

distinction developed in the first chapter. I argue that opponents to simulation theory have failed to distinguish three different objections each of which deserves its own discussion. The first is the question of whether sensorimotor representations could be sufficient to apply concepts. The second is whether they could be sufficient to acquire concepts. Finally, a different question is whether they could be sufficient to individuate concepts. I argue that all three objections, once separated, can easily be answered, at least in principle, by the simulation view of concepts. However, I also make clear that the objection to simulation theory is ultimately an empirical question that requires further research and should not be ruled out on a priori grounds. This work was published in *Philosophical Psychology*.

In chapter 4, I show how abstract concepts, in particular social kind concepts, could be acquired. I argue that we can acquire abstract social concepts by means of what is perceptually available. I endorse an externalist theory of concept acquisition according to which we can rely on correlations between real kinds and their non-defining superficial properties. This is one of so far very few papers where externalism is applied to social kinds. The most interesting result from this work is that it requires social kinds to be mind-independent and stable in a stronger sense than many social constructionists would like to admit. In particular, I argue that many socially constructed kinds like arguably *men o*r *white* cannot be such that they exist because we conceptualize individuals as white and male. Both kinds must have existed before we thought about anything as a white male. This is a strong realist view of social kinds. This work was published in *Synthese*.

Chapter 1: Concepts and Categorization: Do Philosophers and Psychologists Theorize about Different Things?

Abstract: I discuss Edouard Machery's claim that philosophers and psychologists when using the term 'concept' are "really theorizing about different things" (2009, p. 4). This view is not new (e.g., Rey, 1983, 1985), but it has never been developed or defended in detail. Once spelled out we can see that Machery is right that the psychological literature uses a different notion of concept. However, Machery fails to acknowledge that the two notions are not only compatible but complementary. This fits more with the traditional view according to which philosophers and psychologists are merely interested in different aspects of the same kind (e.g., Peacocke, 1992). The main aim of this paper is then to show how precisely the two notions of 'concept' relate. Distinguishing them resolves the long-standing debate on whether concepts can be prototypes and allows me to formulate success conditions of a theory of categorization that are independent of the success conditions of a theory of concepts.

1 Introduction

Edouard Machery (2009) argues that the uses of the expression 'concept' in philosophy and psychology are so different that philosophers and psychologists "are really theorizing about different things" (Machery, 2009, p. 4). This proposal has received little and mostly negative attention in the literature (see, e.g., the commentary to Machery, 2010a), even though it may have tremendous consequences. For example, if it is true that psychologists use a different notion of concept, the widely discussed philosophical objections to psychological theories of

concepts, including Fodor's (1998; 2004) objections to prototype theory, might be based on a confusion.[3]

The aim of this paper is to develop in detail Machery's claim that philosophers and psychologists use the expression 'concept' in different ways, and also show where exactly he is going too far. I argue that the core of his claim is a distinction between concepts as constituents of thoughts and concepts as the explanans of our higher cognitive abilities, especially categorization. The importance of this distinction has already been pointed out several times in the literature, especially by Georges Rey (1983; 1985; 2010), but has not been developed in detail. Once this distinction is spelled out we can see that the two notions are not only distinct in important respects, but complementary. This is in accordance with the traditional view according to which philosophers and psychologists theorize about different aspects of the same kind (Peacocke, 1992).

The main part of this paper is then devoted to teasing apart the two notions of concept used in philosophy and psychology. In particular, I argue that theories of concepts as constituents of thought and theories of concepts as that which explains higher cognitive behavior have different individuation and success conditions that are usually lumped together (e.g., by Rey, 1983 and Prinz, 2002). This disambiguation not only dissolves the long-standing debate on whether concepts can be prototypes, but also contributes to a better understanding of the requirements of a successful theory of categorization and other forms of higher cognition in psychology.

2 What are concepts?

One way to approach the question of what philosophers and psychologists mean by the expression 'concept' is to review how central texts in the literature introduce the term. In philosophy of psychology, concepts are usually introduced as *that which constitutes mental states*, such as beliefs or desires (Rey, 1985; Fodor, 1998; Peacocke, 1992; Prinz, 2002; Machery, 2009; Nimtz & Langkau, 2010). For example, according to Fodor (1998, 6):

[3] I use capital letters to denote concepts, single quotation marks to denote words and italics for emphasis and to denote properties.

> Very roughly, concepts are constituents of mental states. Thus, for example, believing that cats are animals is a paradigmatic mental state, and the concept ANIMAL is a constituent of the belief that cats are animals.

Similarly, according to Rey (1983, p. 237),

> Concepts seem to be the very stuff of which cognitions are made. At any rate, cognitive states like beliefs and preferences, with which many of us hope to explain behavior, seem to involve relations between agents and, roughly speaking, conceptual contents.

In psychology, the expression is commonly introduced as denoting *that which explains everyday higher cognitive abilities*, such as inference- and analogy-making, especially categorization (Smith, Medin, & Rips, 1984; Murphy, 2002; Machery, 2009;). For example, according to Allen Cruse (1999, p. 130):

> Concepts have the status of categories: they classify experience and give access to knowledge concerning entities which fall into them.

Similarly, according to Barsalou (2012, p. 239):

> Componential knowledge in the conceptual system supports a wide variety of basic cognitive operations, including categorization, inference, the representation of propositions, and the productive creation of novel conceptualizations.

The question that I am concerned with in this paper is how these two notions relate to each other. In other words, how does that which constitutes propositional attitudes relate to that which explains categorization and other higher cognitive behaviors.

2.1 The Received View

According to what Machery calls "the received view", philosophers and psychologists interested in concepts really theorize about the same notion. He further distinguishes between what he calls "the simple account" and "the foundationalist account" as versions of the received view. According to the former, philosophers propose individuation and possession conditions of concepts, while psychologists investigate how and whether these conditions are met.

According to the latter, psychologists ascribe concepts to cognitive systems in order to explain phenomena like categorization and analogy-making, while philosophers provide a psychosemantic theory of how these concepts can be about things in the world, i.e., how they can have content (based on what Lewis, 1970, calls a "foundational theory of meaning").

Machery's main objection to the simple account is that it does not describe the actual practice of psychologists. Psychologists do not usually read the philosophical literature and only then investigate whether and how people meet the conditions philosophers propose. Moreover, he argues that the simple account assumes a relationship of subordination that seems unwarranted. Machery's main objection to the foundationalist view is that the methods philosophers use to propose individuation conditions are deficient. Machery especially criticizes the use of thought experiments by Kripke (1972), Putnam (1973) and Burge (1979), referring to his 2011 study that suggests cultural differences in the evaluation of twin earth cases (see also Machery, 2017).

None of Machery's objections to the two versions of the received view are convincing. First, as Machery (2009) himself acknowledges, the project of understanding how the different uses of 'concept' relate is not just descriptive, but normative. Thus, even though the received view may not describe the actual practice of philosophers and psychologists interested in concepts, it may still describe an ideal one. Secondly, psychologists do sometimes read the philosophical literature on possession conditions. This is the case in particular when it is not clear whether the attribution of a particular concept is justified, as in the case of children (e.g., Carey, 2009) or non-human animals (e.g., Wynne, 2001).

Machery's methodological objection to the foundationalist view is also not convincing. First, from the assumption that the use of thought experiments to investigate metaphysical theories of content is sometimes problematic (e.g., due to cross-cultural differences) it does not follow that such a methodology cannot inform our theories at all. Secondly, a deficient method may still produce the correct results. Thirdly, it is not clear whether proposing thought experiments is the only way we can support a foundational theory of content (see e.g., Yli-Vakkuri, 2018).

Finally, at least for the present purpose, a distinction between two versions of the received view is not necessary because they seem to essentially describe the same relation. According to both accounts, philosophers propose individuation and possession conditions (based on a foundational psychosemantic theory that will be further introduced and discussed below), while

psychologists investigate how these conditions are met. In this paper, I want to defend the received view, but also show in what way it needs to be complemented with what I call the "difference account".

2.2 The Difference Account

According to Machery's difference account, philosophers and psychologists tend to associate different descriptions with the expression 'concept'. Philosophers refer to that which constitutes our beliefs and desires, while psychologists refer to that which explains our higher cognitive abilities, especially categorization. This characterization is in accordance with the way the notion is commonly introduced in the two disciplines and leaves open what kind of entity is picked out by the description.

The little attention that Machery's difference account has received so far is mostly negative. Georges Rey (2010), for instance, objects that if concepts were identical to that which allows us to make everyday categorizations we would not be able to understand why most people hold beliefs that seem to contradict their common categorizations of objects. For example, if our concept of doctor were identical to our prototype of doctor, we could not explain why we usually also categorize people as doctors who do not resemble this prototype.

In addition, Rey (2009) objects that if we did not understand concepts as the constituents of propositional attitudes, we could not explain how we are able to communicate with each other. He argues that we could not, for instance, describe the phenomenon of the Müller-Lyer illusion, unless we presumed that people share the concepts LONGER THAN and SAME LENGTH.

Edwards (2010) agrees that there might be a difference in interest between philosophers and psychologists, thus agreeing with Machery's description of the use of the term 'concept'. He disagrees, however, that this justifies the claim that their notions of concept or even their goals are fundamentally different. According to Edwards, psychologists have been using the notion of concept in a very loose manner, mostly as a kind of mental representation that explains everyday categorization. This, he argues, does not mean that they are talking about different kinds of things.

A similar point is made by Lalumera (2010), who objects that, because of the psychologically real phenomenon of conceptual change, we need to determine the identity of a concept

externally. Lalumera worries that, otherwise, we would not be able to explain how people with different experiences can possess the same concepts. How could a chemist and a child, for example, both have the same concept of water? He concludes that if we were to say that the child and the expert have different concepts of water, we would have defeated the very purpose of the notion of concept.

Margolis & Laurence (2010), too, stress that the possibility of psychologists and philosophers having different goals does not necessarily mean that they employ fundamentally different notions. For example, gender might be investigated with different goals in mind, depending on whether one works in the framework of gender studies or whether one is a policy maker. However, this does not justify the claim that both work with different concepts of gender.

None of these authors disagree with the strongest and most charitable reading of Machery's claim, i.e., the difference account. Instead, they seem to presuppose that Machery wants to reduce the philosophical notion of concepts to the psychological one. Furthermore, they seem to assume that they can object to Machery's position if they point out how important the philosophical notion of concept is even in psychology. However, Machery does not deny this (see especially Machery, 2005, p. 446). He also does not argue that the expression 'concept' should be used only for those entities that allow us to categorize our environment. His incentive is to show that the notion of concept differs in philosophy and psychology, which does not commit him to the claim that we should reduce the philosophical notion to what he takes to be the psychological one.

On the other hand, Machery himself might not be fully aware of exactly where the disagreement lies between him and his critics. This would explain why Machery finds Edward's and Rey's objections “puzzling” (2010a, p. 234) and not, for example, based on a misunderstanding. Note that Rey, himself, argued in several of his publications that the study of concepts in psychology is legitimate, as long as it is recognized that what is being researched are not concepts in the philosophical sense but our epistemic access to our concepts (Rey, 2010, 1985, 1983). Machery not only agrees with this – I take it to be his main point.

That the debate is based on a misunderstanding would also explain Machery's (2010b) reaction to Rey's observation that we do not think that doctors always have to look like prototypical doctors. Machery responds that Rey cannot explain our fast categorization abilities if he does

not allow prototypes to play at least some role in our theories of concepts. Similarly, he argues that if we were to identify concepts only with constituents of beliefs that we arrive at after reflection, this would make it much more difficult for psychologists to understand many common stereotypes (Machery, 2010a, p. 232). This reply only makes sense if Machery is thinking of concepts as those entities that explain how we make immediate classifications, while subsuming the content of our reflections under what he calls "background knowledge" (see Machery, 2009; 2015). Rey, however, merely objects to Machery's use of the term 'concept' and not to any specific psychological theory of how we apply our concept of doctor.

Thus, it seems that the way the difference account has been discussed in the literature is based on a misunderstanding. In the next section, I want to develop the difference account further and show why exactly it is justified and important to distinguish between that which constitutes our beliefs and that which explains our higher cognitive processes, such as categorization. Finally, in the last section, I reconcile the difference account and the received view and show why the philosophy and psychology of concepts complement, rather than oppose, each other. In other words, I argue that, in a sense, both the difference and the received views are correct.

3 Concepts and Categorization Devices

To make the distinction between that which constitutes thoughts (beliefs, desires, hopes etc.) and that which explains a person's higher cognitive abilities (categorization, decision-making and so forth) explicit, I call the former "concept" and the latter "categorization device" (also to avoid the problematic term 'conception' that is used by some authors including Rey, but is usually not properly defined). This terminological decision does not reflect any criticism I have towards psychologists' use of the expression 'concept'. Instead, the aim of this section is simply to tease apart the two notions in the clearest way possible.

3.1 The Individuation and Possession Conditions are Different

3.1.1 Individuating Concepts

Concepts (as the constituents of propositional attitudes) are, at least partly, individuated by what one might call their "semantic content" (i.e., their reference). For example, we know that DOG is the concept of dog because it is about dogs. It would not be the concept of dog if it was about

typical dogs, in which case it would be called TYPICAL DOG. It is debated whether semantic content is sufficient to individuate concepts. Fodor (1998), for instance, argues that concepts are also individuated by their so-called "formal properties", while other philosophers, most famously Frege (1948), thought we also need to take into account what one might call a concept's "epistemic content".

One reason why reference or semantic content is not enough to individuate concepts is that people can hold contradictory beliefs about the same reference. For instance, I can believe that Hesperus is a planet but deny that Phosphorus is a planet even though both have the same reference, namely the planet Venus. Thus, if 'concept' refers to that which constitutes our beliefs then we may need an account of conceptual content that cuts concepts into smaller pieces.

A related but independent issue central in the philosophical literature is the more foundational question of what facts make it the case that the respective concept has the semantic content it has. For instance, what makes it the case that the concept of bachelor is about bachelors and not about dogs or typical bachelors? An answer to this question determines the answer to another important question, namely the question of what it takes to possess a concept, i.e., its possession conditions. This question is relevant both for philosophers and psychologists because it determines the point at which it can be said that a concept has been acquired or whether e.g., a non-human animal has them.

The concept literature usually distinguishes between internalist and externalist foundational theories of content. According to the former, the semantic content of a concept is determined by a set of represented descriptions or inferences that pick out the respective reference. An example of such a set of representations or inferences in the case of BACHELOR is that bachelors are unmarried and male. Consequently, if BACHELOR refers to bachelors by means of descriptions or sets of inferences, an individual can think about bachelors if and only if they know the relevant description or are able to draw the relevant inferences.

According to externalist theories, the content of BACHELOR is determined by a real social or biological property in the world and the right causal historical relation between the individual and the respective property or extension. Different versions of externalist theories of mental content have been proposed (e.g., Millikan, 1998; Fodor, 1998) and the details cannot be

discussed here. For the present purpose, however, it suffices that the main difference between semantic externalists and internalists is that the former argue that our concepts can refer even if our psychological states (e.g., our beliefs) are deficient, in the sense that they could not pick out the relevant property or individual. Consequently, according to many externalists, I can possess the concept of bachelor even if, for instance, I lack the correct beliefs about the nature of bachelors or am unable to draw the right inferences, as long as I stand mentally in the right, e.g., causal, relation with actual occurrences of bachelors or the social kind of bachelor.

The question of whether one is an internalist or externalist about mental content not only determines the possession conditions of a concept, but also influences how one individuates concepts. An externalist like Fodor, for instance, can assume that concepts are individuated primarily by their semantic content, even if we sometimes need to posit two formally distinct concepts with the same semantic content (Fodor, 1998). Internalists, on the other hand, need to show how semantic content can be determined in a more internalist way that does not rely as strongly on the external world. Thus, internalists usually argue that our concepts are identical to a set of more basic concepts. In the case of the concept of bachelor this would arguably be the complex concept UNMARRIED ADULT MAN.

3.1.2 Individuating Categorization Devices

Concepts in psychology, i.e., that which explains categorization and other cognitive inductive abilities (what I call "categorization devices") are individuated differently. The main methods psychologists use to determine what information an individual typically uses to make certain categorizations, e.g., that dogs typically bark or that robins are birds, are property verification and property naming tasks (see Murphy, 2002 or Machery, 2009, for summaries). For instance, in order to determine which information an individual uses to apply the concept of bird, many studies simply ask participants to name properties of birds or to categorize different animals *as* birds. This research has been extremely fruitful and produced, for example, the finding that typical birds are recognized faster than atypical birds, suggesting that our categorization device of BIRD processes information about the typical features of birds faster (Rosch, 1983).

A recent debate in the philosophy of psychology has begun to look at the notion of categorization device more closely. It is being discussed, for instance, whether we should reserve the notion of categorization device (what in this debate is called "concept") only for information that is immediately retrieved independently of context. I call this view

"categorization invariantism" (Machery, 2009; Mazzone & Lalumera, 2010).[4] Furthermore, at least Machery (2009) argues that two sets of conceptual mental representations constitute two distinct categorization devices if a) both sets lead to contradictory beliefs, whereby the retrieval of the one set "defeats" the retrieval of the other and b) one set does not immediately cause the retrieval of the other.

Machery's (2009) example to illustrate these criteria is the concept of tomato. Botanically speaking, tomatoes are fruit, even though many or most people consider them to be vegetables. Machery's aim is to describe the behavior of subjects who know that tomatoes are fruit and yet frequently, by default so to say, classify them as vegetables (for example in the context of grocery shopping). Machery's proposal is that these people have (using my terminology) two categorization devices of the same concept of tomato. They have two categorization devices associated with the same semantic content because they have two contradictory sets of default beliefs of tomatoes, namely that they are both vegetables and fruit, and because it is at least plausible that whenever they form the belief that tomatoes are vegetables they do not immediately retrieve the knowledge that they are fruit.

What I call "categorization contextualists" (Barsalou, 1999; Prinz, 2002; Löhr, 2017) reject the regimentation of the notion of categorization device (what they call "concept") that is demanded by categorization invarianists. Instead, they argue that the set of beliefs that constitute our categorization devices need not be stable, but can change depending on context. A minimal condition for being a categorization device is thus not being context independent and retrieved by default, but being able to explain the inductive behavior of the individual in a given context. For instance, in the context of a supermarket the same categorization device might consist of the information that a tomato is a vegetable, while in a different context it might consist of the belief that it is a fruit.

3.1.3 Categorization Devices and Epistemic Content

In what way, exactly, do the individuation conditions of concepts and categorization devices differ? For instance, an immediate objection to this claim might be that at least Machery's

[4] I call this view "categorization invariantism" in order to distinguish it from a different theory in philosophy of language called "semantic minimalism" or "invariantism" (e.g., Borg, 2012). Similarly, I speak of "categorization contextualism" in order to distinguish contextualists about categorization devices like Barsalou (1999) or Prinz (2002) from semantic contextualists like Travis (2008) or Recanati (2010).

criteria for individuation strongly resemble the ones for epistemic content mentioned above. At least their motivation appears to be the same. Machery needs to posit two categorization devices in order to explain contradictory beliefs about tomatoes. Similarly, Frege needed to posit two concepts of Venus in order to explain how we can find it interesting that Phosphorus and Hesperus are identical. So, it looks as if what I call "categorization device", and what Machery, Barsalou and others call "concept", might just be the same notion of concept whose instances are individuated not only by means of their reference but also by means of their epistemic content.

The question of how we individuate the epistemic content of a concept, e.g., of BACHELOR, is complex and cannot be fully discussed here. However, at least traditionally it is assumed that if we can hold two contradictory beliefs about the same reference we have to do with at least two different epistemic contents and consequently with at least two different concepts. This epistemic and semantic individuation of the concept is compatible with different views on how to individuate its corresponding categorization device. According to categorization invariantists (e.g., Machery, 2009), for example, the same semantically and epistemically individuated concept could be associated with several distinct stable categorization devices that may consist of information that is very distinct from the concept's epistemic content, e.g., in the case of BACHELOR that bachelors are typically young men who live alone.

Similarly, according to categorization contextualists (Barsalou, 1999; Prinz, 2002; Löhr, 2017), a categorization device can be more flexible than a concept's epistemic content and may consist simply of the beliefs that are relevant in the respective situation. In the case of BACHELOR, for instance, my corresponding categorization device might thus consist of UNMARRIED ADULT MALE in one context and of YOUNG PERSON WHO LIVES ALONE in another. What matters for the notion of categorization device and the study of how people commonly apply their concepts is thus not to capture the actual epistemic content of a concept, but what explains the actual categorization behavior of the individual in different contexts. Moreover, while it is, at least in principle, admissible for our categorization devices to change their constituting beliefs in a context-sensitive manner, it would be at least much more controversial to claim that the content of a concept is dependent on the context.

This does not mean that 'categorization device' is not an epistemic notion or that the epistemic content of a concept and that which we use to apply a concept could not turn out to be identical.

This would be the case, for instance, if the epistemic content of a concept (e.g., that bachelors are unmarried and male) is in fact usually used to apply this concept. However, this should not be presupposed and is often not the case. For example, if it was found that most people apply BACHELOR as soon as they learn that their young male neighbor lives alone it would not follow that the belief that bachelors typically live alone is therefore the epistemic content of the corresponding concept (although one might argue for such a position, of course). Moreover, it would unnecessarily constrain the psychological research on how we in fact apply concepts if that which we use to make categorizations had to be identical with its content.

Secondly, it is usually assumed that epistemic contents determine the reference of a concept and contribute to the truth conditions of a sentence.[5] However, there is no good reason why we should take for granted that that which explains how we actually apply our concepts, should also have to determine the reference of a concept. While the prototype of bachelor[6] may, as an empirical fact, be what we commonly use to apply the concept of bachelor to individuals, it would at least require good arguments to make a convincing case that this prototype could play the role of the epistemic content of the concept of bachelor if, strictly speaking, it does not refer to bachelors, but to typical bachelors.[7]

Another way to make explicit why 'concept' and 'categorization device' are fundamentally different notions is that it is possible that one can have an epistemically and semantically individuated concept without having the appropriate categorization device. Imagine, for instance, that internalists are right that in order to possess the concept of water I need to know that water is essentially H2O. It is conceivable that I possess WATER (i.e., I know that water is essentially H2O) but lack the typical categorization device of WATER if, for instance, I have no idea what water typically looks like.

Similarly, we can imagine cases in which I have the typical categorization devices of a concept without possessing the corresponding epistemically and semantically individuated concept. If, for instance, I possess the prototype of water, but do not know that water is essentially H2O

[5] However, what is stored in a categorization device may contribute to or even explain the *intuitive* truth conditions of a sentence (see Del Pinal, 2016 for what I take to be a similar claim).
[6] In the following I use a deliberately simplified version of prototype theory. I want to show that even this simple version (e.g., Rosch, 1983) is immune to the common objections often raised against it. Updated versions can be found in Hampton (2000, 2006), Rosch (2011) or Del Pinal (2016).
[7] Although such an argument can be made of course. Del Pinal (2015), for instance, argues that epistemic content (what he calls c-structure) does not need to determine a concepts' reference.

(again assuming that this knowledge is part of a concept's possession conditions) I will, consequently, make the same categorizations as most people in my community (because I possess the same categorization devices), but lack the concept of water and the ability to think about water (because I do not meet the relevant possession conditions).

It may thus seem as if categorization devices resembled more what Frege called *Vorstellungen* or *ideas*, i.e., more or less subjective intuitions and feelings one associates with a concept, rather than the epistemic content of a concept. However, this is only seemingly the case. Since the notion of categorization device is much more constrained than the notion of subjective associations and feelings, I argue that what I call "categorization device" and what many psychologists call "concept" denote a fundamentally different kind of representation that is neither reducible to epistemic content nor the rather unconstrained notion of associations. I submit that systematizing this kind of representation for the first time and starting a theoretical debate on its individuation conditions is one of Machery's (2009) main contributions to the concept literature.

3.1.4 Combining Theories of Concepts with Theories of Categorization Devices

Another reason why 'concept' and 'categorization device' express distinct notions is that we can combine any theory of concepts with any theory of categorization devices. Especially this point has often been misunderstood and ignored in the literature. For example, semantic internalism, the view that conceptual content is determined by a set of internal epistemic states, is sometimes considered to have the advantage that it offers a relatively straightforward explanation of how we apply and acquire concepts (see e.g., Laurence & Margolis, 1999; Prinz, 2002 or Michael, 2015 for this claim). That this advantage is illusory can be seen once we distinguish that which determines the content of a concept from that which we use to apply this concept.

Like semantic externalists, semantic internalists need an additional account of how we apply and acquire concepts to objects, just as proponents of causal-historical semantics need an additional account to explain categorization and concept acquisition. As argued above, knowing that HESPERUS is about the evening star or that the content of BACHELOR is UNMARRIED ADULT MALE does not necessarily explain how we actually apply or even acquire these concepts. For the latter, we need empirical research in psychology, i.e., an account of the categorization devices of HESPERUS and BACHELOR.

To give such an account of concept application, semantic internalism can be combined with any theory of categorization, say prototype theory, definition theory or exemplar theory. Imagine, for example, that in order to have beliefs about bachelors we had to know that bachelors are unmarried adult men. This does not rule out that we, nonetheless, identify bachelors by means of beliefs about the typical features of bachelors, e.g., that bachelors are typically young men that live alone. What I call "content definitionism" (as proposed by Peacocke, 1992) can thus be combined with a what I call "psychological prototype theory" (as proposed by Rosch, 1983) just as it can be combined with "psychological definitionism" (what Margolis & Laurence, 2014 call "the classical theory").

We can even endorse semantic internalism and still endorse a causal view of categorization device. Imagine, for example, a detection device (a bachelor detection app on our phones for example) that allows us to track bachelors even without knowing the necessary or typical properties of bachelors. Such an app is conceivable. We can even imagine a device that is extremely accurate. For example, we can imagine that scientists found a specific bachelor gene and that our detection app would always notify us when we are near a person with such a gene, in a lawful fashion. In this case we may still need to hold certain beliefs in order to refer to bachelors, as internalists argue, but we would, nonetheless, be able to recognize them without this knowledge.

Finally, we can imagine that concepts are best individuated without reference to any epistemic states (such as beliefs about typical properties), but primarily in terms of causal relations between concept and referent (as argued, e.g., by Fodor, 1998). This would mean that we are able to form beliefs about water only if we are in the right causal relation with H2O, even if we lack the knowledge that can sufficiently individuate it. This externalist account is compatible with the view that we, in fact, recognize instances as water by means of its prototypical features, i.e., epistemic states according to which water is typically a transparent drinkable liquid. It is even compatible with the classical view (definition theory) according to which we in fact classify objects by means of representing necessary and jointly sufficient conditions.

Fodor's causal historical view of content can thus be combined with a psychological prototype theory or the classical theory of categorization devices without any compromise. Prototype and definition theories can even help to explain how we establish and sustain this connection, as I

argue below, i.e., how we typically acquire concepts (see also Fodor, 1998; Margolis, 1998). Therefore, prototypes and definitions can be categorization devices even if Fodor is right that they cannot be concepts. This, however, should not worry psychologists considering that prototype theory was not developed to give a semantic account of concepts but, first of all, to explain typicality effects in people's categorization behavior (Rosch, 1983).

3.2 The Requirements for a Successful Theory of Each Notion are Different

Philosophers have proposed several conditions that any successful theory of concepts has to meet. Psychologists have usually simply adopted these conditions without questioning whether they actually apply to a theory of categorization and other inductive abilities (e.g., Hampton, 2000; Prinz, 2012; Barsalou, 2012). In the following, I focus on the two most important desiderata for any theory of concepts, namely "stability", the idea that a concept needs to be stable across different circumstances and people, and "compositionality", the idea that any theory of concepts has to explain how concepts can systematically combine.[8] I argue that theories of concepts and theories of categorization devices have fundamentally different success conditions. Categorization devices, too, need to be stable and compose, but in a fundamentally different and much more relaxed way.

3.2.1 Content and Categorization Stability

Stability is the most important requirement for a successful theory both of concepts and of categorization. However, there are important differences. In order for any theory of concepts (as the constituents of thought) to be a theory of concepts at all, it needs to explain how conceptual content can be stable across different circumstances. Any theory of the concept of dog, for example, from which it follows that DOG sometimes refers to dogs and other times to cats (e.g., depending on whether I am indoors or outdoors) is prima facie not a theory of the

[8] I focus on a very strict understanding of these requirements (as demanded by Fodor, 1998 for instance) in order to show that even they are compatible with the more relaxed requirements of a theory of categorization devices. I do not want to rule out that the many attempts to show how context-dependence is compatible with the idea of compositionality or to weaken the notion of compositionality (e.g., by Hampton & Jönsson, 2012, Del Pinal, 2015 or Recanati, 2010) have been successful. However, I would like to add that most of these attempts aim to explain the compositionality of linguistic expressions and not necessarily of concepts. Moreover, at least Del Pinal seems to discuss the compositionality of his notion of a c-structure, which does not determine the reference of an expression. Arguably this notion is very similar to my notion of categorization device. So, one might argue that the compositionality of the c-structure of an expression only needs to meet the success conditions for a theory of categorization devices and not of concepts.

concept of dog, but of dogs when inside and cats when outside (an exception may be indexicals and other essentially context dependent concepts).

Moreover, the stability of conceptual content is crucial for an explanation of the most fundamental property of cognition: rational inference-making. Imagine for example that I believe that Ana is a doctor and I also believe that Bob is a doctor. My inference that both Ana and Bob are doctors can only be valid if the concept of doctor is the same in both premises. In other words, our concepts must be context-independent in the sense that their contents remain the same in different contexts (e.g., when used in different beliefs) in order to support logical inference-making and consequently rationality.

It is also traditionally assumed that concepts need to be stable in order to explain how they can be combined (e.g., Fodor, 1998, but see Del Pinal, 2015). Compositionality is traditionally assumed to be required for an explanation of how we can form new thoughts in systematic ways despite our limited mental capacities (Fodor, 1989; Peacocke, 1992; but see Werning, 2005). For example, if I can believe that Bob loves Ana, I should also be able to believe (at least in principle) that Ana loves Bob. This, according to Fodor, can only be explained if the concept of love is the same in both contexts. If the way the three concepts are put together would change the content of the respective concepts, this would make the systematicity and productivity of thought mysterious phenomena (Fodor, 1998).

Finally, conceptual stability has been considered crucial in order to show how concepts can be shared – a desideratum that is sometimes called "publicity" (Rey, 1983; Peacocke, 1992; Fodor, 1998; Prinz, 2002). More than one person should be able to have the concept of dog, otherwise it is not clear how a theory of concepts can give us a theory of communication. How, for example, could we ever be able to communicate about dogs or even have the same beliefs about dogs? More importantly, how could we ever debate about the truth of a sentence if our thoughts were not about the same kinds of things, i.e., how could there be real disagreement?[9]

Categorization devices also need to be stable, but in a fundamentally different and much more relaxed way. While a theory of concepts arguably needs to show how the same concept can be about the same kind across different circumstances, a theory of categorization needs to show how our categorization devices are stable in the way that explains our actual behavior towards

[9] This constraint, too, has been challenged for instance by Prinz (2012).

perceived regularities in the world. For instance, while our concept of water arguably needs to be about water in all circumstances that are possible in our actual world, our categorization device of water needs to be stable only in the sense that it reflects the perceived or assumed commonalities between instances of water, which can then be used to categorize a substance as water. In other words, the reasons that speak against what I call "content contextualism" (see footnote 3 and next subsection) are different from the mostly empirical and methodological reasons that, arguably, speak against categorization contextualism.

Secondly, a theory of categorization devices can be more relaxed about providing a theory of rational inference making. We can think that if Ana is a doctor and Bob is doctor, both Ana and Bob are doctors and still think about Ana and Bob as doctors in epistemically different ways. For instance, many people are biased with respect to gender and may categorize Ana and Bob as a doctor based on prejudice. Someone might, for instance, only believe that Ana is a doctor when they see her university degree, while their standards for believing that Bob is a doctor might be much lower. A theory of categorization has to account for this behavioral difference in a way that a theory of conceptual content does not necessarily.

This also means that a theory of categorization need not necessarily meet the (often defended but admittedly still controversial) strict conditions for a theory of what I call "content publicity". Instead, theories of categorization only need to meet the more relaxed conditions for a theory of what I call "categorization publicity". For instance, we can assume that two people can talk about doctors, i.e., have the concept of doctor, without necessarily requiring that they must have the same beliefs about what their typical properties are, e.g., what doctors typically look like. We can thus imagine two people arguing about whether Bob is a doctor because of his young age. Such an argument is only possible (at least according to the requirement of content publicity) if both argue about the same individual and property.

Finally, there is an important difference between theories of concepts and theories of categorization devices with respect to flexibility. While concepts arguably need to retain their content independent of the context, i.e., they have to be context independent, categorization devices need to be flexible in order to explain our flexible behavior in a changing and complex world. For instance, even recognizing an object as a tomato is extremely difficult considering all the different colors and shapes that a tomato can have (an old tomato looks very different from a young tomato for instance). Moreover, the more abstract concepts are, the more flexible

our categorization devices need to be. Consider, for instance, how difficult it is to find a commonality between different vegetables that also explains how we recognize objects as vegetables, or how difficult it is to capture the different methods we use to recognize something as a piece of art in a single stable categorization device (Löhr, 2017).

As argued above, categorization invariantists and categorization contextualists propose different answers to this, in my view, categorization-specific desideratum of flexibility. The former argue that we may have several different categorization devices for the same concept that we retrieve in a context-dependent manner. In the case of TOMATO this would mean that we have at least two categorization devices, which are retrieved flexibly depending on context. The latter typically argue that the information stored in a categorization device can change depending on the context. Again, this would allow an attribution of just one categorization device per concept, which, however, changes its constitutive beliefs or information flexibly depending on the demands of the context.

What I call "categorization stability" is thus a much less demanding notion of stability than the one of content stability. It is less demanding because it need not account for any of the explananda that we need concepts to be stable for, such as sameness of content, rational inference making and content publicity. Moreover, categorization devices can be and should be context dependent (to a degree at least) and need not be shared. Finally, categorization devices also need to account for the flexibility of our behavior, which is a desideratum that does not necessarily apply to a theory of concepts.

3.2.2 Content and Categorization Compositionality

Another commonly proposed desideratum for any theory of concept is what I call "content compositionality". As mentioned above, it has often been argued that one of the main reasons for postulating concepts in the first place is to explain what I call the "content productivity" and "content systematicity" of thought (Fodor, 1989). Content productivity is the phenomenon that we can form an unlimited number of new beliefs despite our limited mental capacities. It has been traditionally assumed, probably falsely (Werning, 2005; Pagin, 2012), that in order to explain the productivity of thought, concepts must combine in systematic ways (Rey, 1983; Fodor, 1998).

The reason why we need categorization devices to compose is different. We do not posit what I call "categorization compositionality" in order to explain content productivity, but to explain what I call "categorization productivity", i.e., the observation that when I learn about a new category I typically first rely on my current beliefs in order to guess what the new category *could be*, as opposed to, *is* about (i.e., how the respective concept could be applied). This guess need not be correct. While the concept of wooden spoon should be about wooden spoons, my categorization device of the same concept may not allow me to make the correct classifications.[10] So even if I know how to apply WOODEN and SPOON I might still be mistaken when it comes to applying the combination of both concepts. If my guess is incorrect I will either use contextual cues or make further guesses (Machery & Lederer, 2012). In other words, while content compositionality refers to a property of our beliefs, categorization compositionality refers to our ability to form hypotheses about unknown categories.

Imagine, for instance, hearing the word 'pet fish' for the first time. Imagine you know when to apply the concept of pet and the concept of fish, but you have no idea when to apply the concept of pet fish. To explain how we, nonetheless, usually have at least a rough idea of what pet fish are (as opposed to what the actual content of PET FISH is), we require a theory of how we can combine our knowledge of pet and fish flexibly. In the case of PET FISH we would probably, as a first guess, combine the typical features of pets and the typical features of fish to something like a hypothesis of what pet fish could be. At first, we will, perhaps, think of something very large and hairy that lives in the ocean, but for some reason also in the living room of our friend. Then we try to integrate our categorization device of pet and fish in a way that is more appropriate in the given context and reach the conclusion that pet fish are probably small fish that are kept in tanks in the living room (see Hampton & Jönsson, 2012 for a more detailed account of how this could work).

Referring to context in order to explain how we use our beliefs to make sense of new words and categories is legitimate as long as this is in accordance with our actual behavior. Referring to context to explain content compositionality is much more problematic (but again, according to more recent approaches to compositionality perhaps still acceptable) because this would, arguably, risk violating the requirement of content stability. If the content of our complex concepts would depend on context, i.e., if the content of PET FISH would change depending

[10] Again, this notion of categorization compositionality has little to do with the notion of compositionality in philosophy of language and mind. For the more common notion of compositionality see for instance Szabó (2012).

on the situation in the relevant sense (i.e., not in the sense relevant for debates on indexicals or semantic externalism) this would, at least according to Fodor and others, threaten our explanation of how we can make rational inferences and how we can communicate about the same kinds of things.

Some philosophers, most notably Fodor (1998, 2004), have argued that compositionality is a problem for many psychological theories of concepts, especially prototype theory. The main objection is that the typical properties of fish and the typical properties of pets combined do not produce the typical properties of pet fish. This has been challenged (e.g., Hampton, 2006; Prinz & Clark, 2004; Prinz, 2012; Del Pinal, 2016), but none of these attempts can do without reference to context or background knowledge that could adjust the combination "big hairy cuddly animal that swims" to what pet fish actually are (see Machery & Lederer, 2012). If conceptual content must be stable, this context dependence might thus challenge psychological theories of concepts.

However, once we distinguish between content compositionality and categorization compositionality, we can see that theorists of categorization devices can be more relaxed about the problem of stability and context dependence. They can allow that the typical features of pet and the typical features of fish do not compose in a way that produces the actual semantic or epistemic content of PET FISH. Instead they need a theory of categorization devices to explain how we actually combine our categorization devices of both concepts (which a theory of conceptual compositionality need not necessarily).

While prototype theory is thus problematic as a theory of content compositionality, especially as a kind of inferential role semantics, as Fodor (1998; 2004) argues, it can account well for the phenomenon of everyday categorization compositionality. Again, this is because if we do not know what the application conditions of a complex concept are we usually combine the typical application conditions of its constituents enriched by context. In the case of categorization, reference to context is not problematic, while it is at least controversial whether we can allow context to determine the content of our concepts, as this might risk other important requirements of content stability and content publicity.

In summary, there is a distinction to be made between criteria for a successful semantic theory of concepts and criteria for a successful theory of categorization devices, which are usually

lumped together (e.g., by Fodor, 1998; Rey, 1983 and Prinz, 2002). This conflation of concepts and categorization devices has caused tremendous confusion, including the debate on whether prototypes can be concepts, which turns out to be a philosophically rather uninteresting and ultimately an empirical question (i.e., philosophers who doubt that they can be concepts can be indifferent about whether they are at least part of our categorization devices). Once we distinguish between two uses of the expressions 'stability' and 'compositionality' we can conclude that prototypes can perhaps not be concepts (because they cannot, arguably, account for content compositionality and content stability), but that they might still turn out to be partly responsible for our categorization devices (because categorization compositionality can include context and need not account for the requirements of content publicity and stability).

4 Reconciling the Received View with the Difference Account

In the preceding section, I argued that the notions of concept as the constituents of thought and concept as categorization devices differ in important respects. This supports the difference account introduced above in the sense that it is not only true but justified that researchers interested in the constituents of thought and researchers interested in categorization use different notions of concept. However, from the argument that the expression 'concept' is associated with different notions it does not follow that both notions pick out, ontologically speaking, fundamentally different kinds of things or properties. Furthermore, the difference account does not explain how the two notions relate in a more positive sense, i.e., what they have in common.

In this section, I argue that accounts of constituents of thought and categorization (i.e., concepts and categorization devices) relate in the following way:

1) In cases in which categorization devices contain beliefs[11] (e.g., that dogs typically bark) and concepts constitute beliefs then concepts are the constituents of those categorization devices

[11] It may be that categorization is not explained in terms of beliefs and that categorization devices do not contain them. In this case categorization devices do not consist of concepts, but for instance of non-conceptual perceptual representations. I do not want to rule this out. My argument is merely that often psychologists explain categorization by attributing beliefs to people even if these beliefs are presented in terms of lists of features. Only in such cases is the relation between concepts and categorization devices as argued here.

2) Categorization devices allow us to acquire and apply concepts to objects and gather information about our own concepts

The difference account is thus compatible with the received view, according to which philosophers and psychologists theorize, ontologically speaking, about different aspects of the same kind of thing. These kinds are the constituents of thoughts or what I, following the philosophical literature, called "concepts".

Consider, for example, a typical case of concept application, such as categorizing the transparent liquid in front of you as water. This behavior is commonly explained by means of a belief attribution, e.g., the belief that transparent liquids are typically water. Beliefs, we assumed above, consist of concepts. Hence, the aforementioned belief, i.e., that which allows us to make certain classifications (or the categorization device associated with water), consists of the concepts of transparent and liquid. In other words, the term 'categorization device' denotes sets of beliefs, each of which consists of concepts. This suggests that the received view is correct. Philosophers and psychologists do theorize about different aspects of the same kind of things, namely concepts understood as the constituents of thought (beliefs, desires etc.).

But what exactly are these different aspects? First, by constituting categorization devices, concepts provide the semantic content of our categorization devices. Such an account is required because, in order for categorization devices to aid us in our classification behaviors, they need to be about things in the world. Our categorization device for WATER can only consist of TRANSPARENT and LIQUID and can only help us to recognize something as water if its constituents actually refer to these properties. Secondly, the constituents of categorization devices, i.e., concepts, need to (at least according to Fodor) meet the criteria of content stability and compositionality in order to constitute categorization devices. Imagine, again, that our categorization devices for WATER consist of LIQUID and TRANSPARENT, i.e., the belief that water is typically a transparent liquid. This constitution relation is only possible if our concepts of liquid and transparent are stable and can be combined in the semantic senses of stability and compositionally.

While concepts play an important constitutive role for categorization devices, categorization devices play an important role not only for the application of concepts (e.g., for categorization) but also for their acquisition. To see why, imagine first that externalist (e.g., causal-historical)

theories of content are correct. Categorization devices can explain how we establish and sustain a causal relation to the referents of our concepts (Margolis, 1998; Laurence & Margolis, 2011). According to Fodor (1998), for example, the question of how we establish a nomological relationship between agent and reference is to be answered by psychologists. One such option is to say that I have a stereotype (Putnam, 1973) or a prototype (Rosch, 1983) of what dogs normally look like (what I call "categorization device").

If semantic internalists (e.g., semantic inferential role semanticists) are correct, we need a theory that explains how we acquire the ability to draw the inferences that determine concept possession. Unless they are explicitly told what bachelors are, most children probably learn how to use the word 'bachelor' by hearing it in context, for example by hearing someone being called a bachelor. Our first experience with bachelors will thus mostly be controlled by beliefs that are more based on statistical regularities (e.g., that they typically live alone in small apartments) than on necessary and sufficient conditions that define what we mean by the word 'bachelor'. So, in order to learn the necessary and sufficient conditions for being a bachelor, children probably start out with less precise categorization devices that are more based on typical superficial features than the descriptions that can fix the actual reference of our concepts.

Finally, our categorization devices allow us to gather information about our own concepts. Think of the concept of friendship. We may have a very rich notion of friendship, i.e., we are extremely good at applying the concept in various contexts. However, few of the application conditions we associate with this concept are explicit. In order to understand better what we mean by 'friendship', we can use our categorization device of friendship to apply it in different, perhaps imagined, situations, what philosophers call "thought experiments", in order to get a clearer idea about the intuitive application conditions of FRIENDSHIP. This opens the possibility of a kind of conceptual analysis that need not depend on any theory of content (see e.g., Machery, 2017).

So, the two notions of concepts just described, although distinct, are highly related in important respects. This can also explain how the philosophy and the psychology of concepts relate. Philosophers give individuation and possession conditions of concepts, while psychologists investigate how we meet these conditions (as suggested by the received view). However, psychologists also need to show how we apply these concepts to things in the world by means of our categorization devices (as suggested by the difference account). Therefore, the accounts

of how the philosophy and psychology of concepts relate that were introduced above are not incompatible. Instead, they complement each other.

5 Conclusion

The term 'concept' is commonly introduced in philosophy to denote the constituents of our thoughts. In psychology, the term is usually introduced to denote that which explains higher cognition, especially categorization. The difference in use of the term suggests that philosophers and psychologists theorize about different phenomena by means of different notions, namely what I call “concepts” and “categorization devices”. The purpose of this paper was to show in what ways both notions relate and why they have to be distinguished in order to avoid confusion in the interdisciplinary literature. However, I also showed that they are linked in important respects. For example, concepts may constitute categorization devices and we use categorization devices to acquire concepts.

A major advantage of the distinction defended here is that we can now resolve the problem of whether concepts can be prototypes. Prototypes have been thought to be unfit to be concepts because they arguably cannot account for stability and compositionality (Fodor, 1998; 2004). I showed that we need to distinguish semantic and more psychological notions of both stability and compositionality and that the latter are less demanding than the former. However, the objections to prototype theory, especially by Fodor, have focused only on the former notions of what I called “content stability” and “content compositionality”. As soon as the distinction between concepts and categorization devices and the different success conditions of a theory of each are understood in the way discussed here, I predict that other debates in the philosophy of psychology will find similar solutions.

Chapter 2: Abstract Concepts, Compositionality and the Contextualism-Invariantism Debate

Abstract: Invariantists argue that the notion of concept in psychology should be reserved for

knowledge that is retrieved in a context-insensitive manner. Contextualists argue that concepts are to be understood in terms of context-sensitive ad hoc constructions. The recent literature views this issue as ultimately empirical. I review the central empirical studies and show that their conclusions are based on a common mischaracterization of the relevant theories. When the difference between Contextualism and Invariantism is properly understood, it becomes apparent that empirical evidence will not be decisive. The issue, I argue, is not empirical, but purely theoretical. Consequently, the debate should return to theoretical arguments. I offer one such argument: Invariantism fails to account for two important theoretical conditions for any theory of concepts – scope and compositionality.

1 Introduction

One of the main challenges in psychology is to show how cognitive behavior can be regular, reliable and immediate in a world that is complex and constantly changing. The mainstream view in cognitive science is that this is possible because we mentally represent the regularities and commonalities that we find in the world. Weapons do not suddenly turn into flowers and most animals we encounter have to eat, breath and sleep. The technical term for these mental representations in psychology is *concept* (Barsalou, 2012; Machery, 2009).

Contextualism and *Invariantism* are theories about how this notion of concept in psychology is best understood. According to Invariantism, concepts are best understood as context-independent mental representations in long-term memory (Machery, 2015; Barsalou, 2012; Mazzone & Lalumera, 2010; Dove, 2009; Laurence & Margolis, 2002; Keil, 1989). Concepts are thus not *constructed* based on situational demands, but *retrieved* from memory as context-independent, stable entities. This view does not deny that concepts may change over time, but that this change is more long-term.

According to *Contextualism* (e.g., Yee & Thompson-Schill, 2016; Casasanto and Lupyan, 2015; Lebois et al., 2015; Kiefer and Pulvermüller, 2012; Hoenig et al. 2008; Kiefer, 2005; Barsalou 1992, 1987) concepts are best understood as context-dependent, unstable entities. This means that the information that is stored in a concept constantly changes with context, and that most complex concepts do not have a stable set of constituents. To support this claim, Contextualists mostly rely on research showing that our higher cognitive behaviors, such as object-recognition and inference-making, highly depend on context.

The supposed theoretical advantage of context-independent representations is that they give a straightforward account of the intra- and interpersonal stability of thought and communication, including its apparent automaticity and immediacy. For example, when seeing a tiger in the wild, it seems that we do not have to retrieve all of our knowledge associated with tigers to construct a concept on which we base our decision to run away. Instead, we seem to just see a tiger and then immediately react appropriately because our concept of tiger partly consists of the property 'highly dangerous'.

Contextualists, on the other hand, seem to be especially well-equipped to explain how we are able to deal with the complexity of our world and how we can respond appropriately to new or atypical situations or exemplars. For instance, many properties that we associate with tigers might change with context. Not all tigers have stripes and are dangerous, but we are still able to identify them as tigers and make the right inferences. Contextualist can explain this by arguing that we retrieve different representations of features of tigers depending on what kind of an instance of this category we actually encounter.

Both parties usually take for granted that the debate is ultimately decided on empirical grounds, which will be summarized below (Bloch-Mullins, 2015; Machery, 2015; Prinz, 2002; Barsalou, 1987). However, despite a wealth of findings over the past thirty years, no consensus has been reached. I argue that the reason for this is not a lack of thorough research, but that any relevant evidence can be explained by both views. Further theoretical problems, such as the difficulty to define the notion of context, add to the suspicion that a decision between Contextualism and Invariantism based on empirical findings is highly unlikely.

Instead of focusing on empirical evidence, I propose that the debate should return to theoretical arguments based on criteria that a successful theory of concepts has to meet (see for example criteria proposed by Prinz, 2002 or Fodor, 1998). Three such criteria are widely agreed on as necessary for a successful theory of concepts, both in philosophy and psychology (see Machery, 2009; Fodor, 1998 and Rey, 1983 for this distinction). The first criterion is that any theory of concepts ought to explain both predictable and flexible behavior. It is commonly thought that an answer to this question of *stability* will also provide a theory of how concepts can be shared among different people, which appears to be a requirement for communication.

A second crucial desideratum is that any theory of concepts should explain at least a majority

of concepts. A theory that can only account for few concepts, say only concrete concepts (concepts that refer to physical objects), is less successful than a theory that can explain all or most concepts, as long as both are equally explanatorily successful in the other respects. Finally, a theory of concepts ought to explain how simple concepts, such as PET and FISH, can combine into more complex concepts as PET FISH. This *compositionality* desideratum appears to be crucial to explain two other important desiderata for any theory of concepts, namely the *productivity* and *systematicity* of thought.[12]

I propose that even if empirical arguments have failed to yield a decision on the debate of whether we should be Contextualists or Invariantists, we can turn to the just mentioned theoretical considerations. In particular, I argue that while Contextualism can account for stability, Invariantism lacks a convincing account of abstract and composed concepts. I discuss each theoretical desideratum respectively (first stability, then scope and finally compositionality) and conclude that we should give up Invariantism in favor of Contextualism, because although it accounts for stability it lacks an account of the other two.

2 Stability

2.1 Theoretical Arguments

As introduced, the notion of concept as a theoretical construct is posited to explain how we can make sense of a complex and changing environment. Because of its role as an explanans, a concept has to meet certain criteria. The most important desideratum for any theory of concepts is that it has to account for inter- and intra-personal *stability*. In other words, without an explanation of how we detect regularities and commonalities in our environment there is no theory of concepts (Fodor, 1975, 1998; Evans, 1982; Rey, 1986; Laurence and Margolis, 1999; Prinz, 2002).

Invariantists have a straightforward solution to the stability desideratum. They argue that only representations of the most informative properties of all the properties that we might attribute to an entity are to be called its corresponding concept. According to Definition Theorists the most informative properties are the ones that are common to all members of the category.

12 Productivity refers to the fact that we can think unlimited thoughts with a limited set of concepts. Systematicity names the observation that these thoughts can be combined in a logical and systematic matter. For example, it explains how we can form the thought that John loves Mary, but also form the thought that Mary loves John (Evans, 1982).

According to Prototypes Theorists, the most informative properties are the ones that are typical for the respective category. Because they are common to all, or all typical contexts, they can explain our typical behavior, i.e., the behavior that is stable across different context.

Contextualists object that the world is too flexible and complex for these context-independent concepts to explain even typical higher cognitive behavior. They claim that especially unusual and new situations pose a serious challenge for this view. In support of this, they usually rely on empirical evidence that suggests that no single context-independent mental representation is powerful enough to provide the explanation psychologists are looking for. Instead, depending on the task and context, the information or represented properties that are retrieved seem to vary. According to Contextualists, concepts are thus not comprised of the properties that are always or typically retrieved, but the properties that are actually retrieved to determine behavior in the respective situation.

However, Invariantists have found a way to respond to this challenge. They account for unusual situations and atypical exemplars by positing a second kind of knowledge structure called *background knowledge* (see e.g., Machery, 2015). The difference between conceptual and background knowledge is not only that the latter is context-sensitive, but also that it takes longer to retrieve. While conceptual knowledge, according to Invariantists, can be automatic and immediate, context-dependent knowledge requires more computational resources and is thus less automatic and less immediate.

While Invariantists have found ways to explain how behavior can be stable without being rigid, Contextualists too have resources to explain how behavior can be stable. Barsalou (1999), for example, argues that we construct concepts based on the perceptually accessible, multimodal properties and relations of objects. After seeing several instances of the same concept, which most likely activate similar neural states, these instances are represented in so-called *correlated feature maps*. Stability is thus explained in terms of a kind of abstraction mechanism, i.e., the integration of multimodal representations over time. This means that we represent similar objects based on similar experiences as well as our interactions with them. A concept like DOG, for example, is derived from neurons that reliably activate upon seeing certain shapes and colors independent of length, position and orientation of the respective dog. Because these line and vertex detectors are generic, the actual orientation or appearance of the object is abstracted away from.

This account eliminates a common misunderstanding (see e.g., Mazzone & Lalumera, 2010). When Contextualists deny that concepts are stable they neither mean mean that our behavior is not stable nor that our concepts fail to explain this. The Contextualist and Invariantist disagreement is not about whether all of our conceptual representations are stable or not, but about whether one context-insensitive mental entity can explain all the various *typical* situations that we associate with a certain category. Contextualists reject this claim, Invariantists essentially commit to it (Casasanto and Lupyan, 2015; Machery, 2015).

Besides intra-personal stability, a further aspect of stability is inter-personal. If concepts are always changing, as Contextualists argue, how can they be shared? That concepts can be shared is considered a theoretical requirement for any successful theory of communication that operates with concepts (Prinz, 2002; Fodor, 1998; Rey, 1983). Since not all properties that we might associate with a category will be shared, Invariantists argue that only the features that are always or typically retrieved in concept-related situations are shared among individuals. Contextualists, on the other hand, account for inter-personal stability by stressing the similarities between the physical make-up of humans and, again context, e.g., that people from similar backgrounds usually have had very similar experiences (Clark and Prinz, 2004; Barsalou, 1999; Newton, 1996; Tomasello et al., 1993).

This debate cannot be fully assessed here. However, for now it suffices that no consensus on this issue has been reached, and that it has not been shown that context and similar physical make-up cannot sufficiently establish and sustain inter-personal communication. Moreover, it might turn out that the doubts about Contextualists' ability to explain inter-personal stability are exaggerated. According to Fodor (1998), two people can have the same concept (individuated in terms of reference) even if we associate different features or properties with these referents. This finds further support in the contemporary philosophy of language. Proper names like Aristotle, for instance, are said to refer to the same individual even if the descriptions we associate with the name Aristotle turn out to be false. What is important is that a causal connection between the community and the individual referred to is sustained (Kripke, 1972, Putnam, 1973). This theory of reference can and is applied to Contextualism (Barsalou, 1999; Prinz, 2002).

In summary, while Invariantism seems especially appropriate to account for stability,

Contextualism seems especially apt to explain how thought and behavior can be flexible. This does not mean that Invariantists have no theoretical tools to explain flexibility and miscommunication, e.g., by means of background knowledge. However, it also does not show that Contexualists cannot, at least in principle, account for stability in terms of context, causal connection and similarities between speakers.

2.2 Empirical Evidence

It is currently assumed by both parties that because the theoretical arguments have not been able to settle the debate, empirical evidence may be more promising (Machery, 2015; Bloch-Mullins, 2015; Oosterwijk et. al., 2015; Barsalou, 2003; Smith and Samuelson, 1997). Here I review only the main findings that have been central in the recent debate. The aim is not to be comprehensive, but to expose a commonality that I take as highly problematic for both past and future research on Contextualism and Invariantism. In particular I argue that what all the empirical findings have in common is that they show typicality, redundancy and context effects that can all be explained by both positions and therefore fail to be decisive.

One of the first pieces of evidence mentioned in the literature comes from simple property naming and property verification tasks that suggest low *between-subjects*- and *within-subject-reliability* (Barsalou, 1987). In other words, it seems that different participants do not always associate the same properties with the same entity. Moreover, even the same subject may associate different properties with the same referent if asked in different contexts. For example, as Barsalou & Sewell (1984) report, when subjects were asked to name properties they associate with a given category, the correlation between subjects was only around .45.

Similarly, Barclay et al. (1974) (as reported by Machery, 2009, pp. 23) show that when the expression 'piano' is used in a musical context, participants retrieve properties that are related to music. However, when the same expression is presented in a different context, say next to the sentence “your friends are moving out”, participants are more likely to retrieve representations of a piano's appearance, e.g., their weight and shape. That there is little overlap between the retrieved features in both contexts was taken as evidence that concepts are highly unstable and do not just store features common to all members of the group (e.g., Barsalou, 1987).

One problem with these findings is that they could easily be explained in terms of background knowledge. The reason why subjects mentioned different features depending on the context could simply be that much of the retrieved knowledge is not part of the respective concept. Furthemore, Machery (2009) objects that the variation of features between days, reported by Barsalou (1993), was not significant and is actually evidence in favor of Invariantism. The result of one study by Barsalou and colleagues, as reported in Machery (2009, pp. 24), was that seven out of ten features were listed again on a further occasion, which is, according to Machery, evidence that these subjects mostly retrieved default knowledge, adapted to the context by means of background knowledge.

In more recent years, Contextualists have turned their attention to so-called *modality effects*. For example, Yee and colleagues (2016) summarize a finding by Connell and Lynott (2014) that subjects were faster at recognizing a word that was presented visually, as opposed to auditorially, if the referents of this word are usually experienced visually (e.g.,'flower'). Similarly, during an auditory lexical task, subjects were faster at recognizing words that are associated with audition (e.g., 'bark'). This, according to Yee et al., suggests that the knowledge that determines and explains the behavior displayed in the experiments seems to partly depend on the demands of the situation.

However, although this finding is suggestive, and although Invariantists clearly make predictions regarding the time of retrieval (conceptual knowledge is meant to be retrieved faster than background knowledge), Invariantists may find this result little surprising. For example, they could argue that modal information (information stored in the modalities) may have facilitated the lexical decision task because it was part of the respective invariant concept. In the case of 'flower', for example, it is likely that visual properties may be part of the stable concept of flower if we typically identify flowers by its visual properties. In the case that visual information is retrieved after auditory information, Invariantists could argue that visual information belongs to background knowledge.

Especially modality effects have been frequently mentioned by both sides. For example, to support Invariantism, Machery (2010b) cites fMRI (functional Magnetic Resonance Imaging) studies by Hoenig et al. (2008) and James and Gauthier (2003) that show that brain areas are automatically activated when we process a concept even if these areas do not appear to be

required for the given task. They observed, for instance, that auditory and motor areas were activated when novel objects (called *greebles*) were presented visually to subjects who had been trained to recognize them by sound and movement. This, according to Machery, should not be the case if knowledge was retrieved in a context-dependent manner, in which case one would expect an increase in activation only in the necessary areas of the brain.

Another important study that Machery (2015) recently mentioned to support Invarianstim was Whitney, McKay, Kellas, and Emerson's (1985) version of the so-called "semantic Stroop test", which tests the extent to which context influences access to lexical meaning. The classic Stroop test (Jaensch, 1929; Stroop, 1935) presents subjects with different color-words (blue, red, yellow) that are printed with sometimes congruent and sometimes incongruent ink. The task is then to report the color of the ink. Correct performance slows down considerably if the color of the sign does not match the meaning of the corresponding word. For example, if presented with the word 'blue' subjects have difficulties reporting the color of the ink that the word is written in if the ink is yellow.

In Whitney et al.'s study, the target word was also printed in different colors that the participants had to name. In this case however, the words were not color-terms but properties of concrete objects and organisms, such as 'nose' or 'eyes'. It has repeatedly been shown that color naming also slows down in semantically atypical contexts (Connell and Lynott, 2009). In the present study, this means that before seeing the word, participants heard a sentence (representing the context). This sentence ended with a word that is semantically related to the property expressed by the stimuli. For example, in the case of the concept of rabbit, the sentence "bees sting" would represent a very dominant feature, while the sentence "bees buzz" would represent a less dominant feature.

The authors found that context influenced conceptual access only for less dominant properties. For example, the context sentence only influenced the retrieval of less dominant features of RABBIT (e.g., that rabbits hop), but not of highly dominant features, as, for instance, that rabbits have fur. Machery (2015) interprets this as evidence against Contextualism because context made no difference to the retrieval of the *highly-dominant* or supposedly context-insensitive property conditions. It only made a difference in the case of atypical or *low-dominant* properties, which can be attributed to background knowledge. Machery concludes that the more dominant properties function as cores, i.e., as invariant features of concepts.

Finally, Machery (2015) reports an experiment by Barsalou (1982) in which participants had to complete a property-verification task. An example of a context-independent property was 'smells' for the category 'skunk', while a context-sensitive property was 'can be walked on' in the case of the category 'roof'. It was assumed that, while not all roofs can be walked on, participants think that all skunks smell. Again, the basic intuition is that some properties are context-sensitive ('can be walked on') and thus belong to a subject's background knowledge, while others ('smells') are more typical and frequent and thus independent of context. This intuition was supported. Subjects indeed verified 'smells' for 'skunk' significantly faster than 'can be walked on' for 'roof'.

None of these studies can seriously challenge Contextualism either. The problem is that they all present some kind of regularity effect, mostly typicality or redundancy effects. This effect is interpreted as showing that we retrieve information not because it is relevant, but simply because it is typical or frequent. As summarized above, Hoenig et al. (2008) found seemingly irrelevant brain areas that were active during concept retrieval. Whitney et al. (1985) found that typical information was retrieved faster even if it was not necessary for the task, and Barsalou's (1982) study showed that context did not interfere with the retrieval of typical features.

It is usually not made explicit why Contextualists should have difficulties with these findings. Just as any reasonable Invariantist view must be allowed to account for some degree of context variance, e.g., by postulating background knowledge, Contextualism must be granted to have at least some degree of freedom to account for some invariance and regularity. Otherwise, Invariantism would be trivially true. First, because concepts have to, by definition, explain stable and reliable behavior. Second, because typicality effects are so robust that they are accepted by most psychologists, including prominent Contextualists as Barsalou (1999) and Prinz (2002).

That Contextualists accept these findings is justified because typicality effects are not sufficient for proving the existence of stable prototypes. Other explanations of these effects might turn out to be more advantageous. Contextualists can explain these and other effects, such as recency and frequency effects, in terms of a simple construction mechanism, such as Hebbian Learning, which gives typical and frequent information a processing advantage. Hebbian Leaning is not necessarily evidence for there being context-independent concepts in the strong sense that

Invariantism requires (see e.g., Pulvermüller, 2013 for such an account).

The Hebbian learning rule is meant to describe how the brain makes new connections between sensorimotor inputs and how it detects and stores correlations in the environment. The rule predicts that the connections that "fire together" become stronger, while experiences that less often occur together are representationally inhibited. Contextualists are not committed to rejecting this proposed mechanism. Even if context suggests constructing an atypical concept of, say, 'nose', Contextualists can leave room for a competition between typical and context appropriate features of a category without giving up their core commitment that concepts can contain context-dependent information.

Finally, it is hardly mentioned in the literature that some knowledge is, for relatively uninteresting reasons, context-insensitive and poses therefore no challenge to Contextualism. Why, for instance, should the result for the property 'smells' in the case of SKUNK be evidence against Contextualism? All I, and most other non-experts, know about how to distinguish a skunk from a similar animal is that it smells bad. Since Barsalou's (1982) study does not suggest otherwise, it is reasonable to assume that the subjects' knowledge about skunks is equally limited. So why assume that the knowledge that skunks smell bad is not part of the subjects' conceptual construction of skunk every time they see the word 'skunk' appear on the laboratory screen even when the context does not require it?

It thus seems that in the case of both Invariantism and Contextualism, the only kind of knowledge that could support either position can easily be accounted for by the opposing view. On the one hand, the evidence that is usually mentioned in support of Contextualism is that the knowledge that is retrieved during a task changes with context. This cannot reasonably be denied by any Invariantist who could object that background knowledge could be responsible for these effects. On the other hand, Invariantists put forth redundancy and typicality effects, which can be explained by limited knowledge and basic neural mechanisms that give typically and frequently retrieved knowledge a processing advantage. Consequently, it seems that both views are too flexible to generate hypotheses that could lead to falsification, which may be the reason why so little consensus has been reached based on empirical evidence.

2.3 Context

A further issue that is often thought to contribute to the difficulty of falsifying both

Contextualism and Invariantism is that, so far, none of the parties have been able to define what exactly they mean by the word 'context'. Machery (2015, pp. 574) suggests that the burden to provide such a definition lies with Contextualists, but this is unjustified as Invariantists also make use of the notion (both in the case of background and context-insensitive knowledge).

Bloch-Mullins (2015, pp. 943) even expressed the concern that since the notion of context has not been defined, Contextualism, too, is not fully specified. However, this criticism is exaggerated. In experiments, for example, the notion is usually operationalized in terms of priming sentences that are presented temporarily prior the to target. So what Contextualists mean by context is not a complete mystery, even if the notion is so far only operationally defined. Nonetheless, Bloch-Mullins is right in so far as sentences can only induce a new context, but not suffice for its individuation. The problem remains that almost anything can be regarded as related to context, unless researchers were to agree on some constraints.

One might think that an agreement on such constraints is unlikely. For example, Mazzone & Lalumera (2010) mention two understandings of context: A narrow notion, according to which context is defined in terms of a more or less objective list of time, environment and agent, and a more expansive one that also includes mental states and more fine-grained descriptions of the environment. It seems that Invariantists could always explain problematic evidence in terms of the more narrow definition, while Contextualists could always refer to the more expansive one. Unfortunately, this is precisely what many Contextualists do (e.g., Casasanto and Lupyan, 2015), which adds to the suspicion that both views are extremely difficult to falsify.

Contrary to these authors, I suggest that the problem of context is hardly what drives the lack of consensus. Instead, it hinges on a more fundamental one, namely the same that was described in the previous subsection. We could for instance imagine a situation in which both parties agreed on the most inclusive notion of concept (i.e., almost anything would count as a new context). This would still leave room for the possibility that stable knowledge was retrieved across obviously different contexts. Imagine further that this overlapping knowledge would explain all of our typical behavior. In this case we would have a definition of context that both parties can agree upon and a powerful account of conceptual knowledge. The more fundamental problem is that Contextualists could again argue that seemingly context-independent knowledge is reducible to typicality, frequency or recency effects that no Contextualist need deny.

Thus, the more fundamental problem seems once more to be that the only evidence that could support Invariantism, i.e., some kind of regularity effect like typicality or frequency, can also be explained by Contextualists. The notion of context is a problem, but it is far less problematic than some authors have claimed and it is not the most fundamental.

2.4 Concept Individuation

Another way that Invariantists have responded to context effects is to cut concepts into smaller pieces. For example, Machery and Seppälä (2011) present linguistic evidence that suggests that subjects often subsume the same concept under different superordinate concepts depending on the context. For example, they show that participants in one context subsume the concept of tomato under the concept of fruit, and in another context under the concept of vegetable (for a counter argument see Zarl & Fum, 2014). This appears to be strong evidence for Contextualism because it is not clear how Invariantism could be maintained when even knowledge about the superordinate class that a concept belongs to depends on a context (this information seems to be a paradigm case of context-insensitive knowledge).

Machery and Seppälä, however, reject this interpretation. They refer to Machery's (2009) conditions of concept individuation, according to which two judgments that are either contradictory or do not immediately facilitate each other's retrieval belong to different concepts. For example, Machery argues that if a category like 'tomato' is associated with two or more contradictory judgments, such as “is a vegetable” and “is a fruit”, it must be represented not by one context-dependent, but by two context-independent concepts.

This way of individuating concepts in psychology is motivated by the following independent grounds and should thus be taken seriously. Recall that the notion of concept in psychology has an explanatory function. Concepts ought to explain how we typically sort objects into categories and make certain typical inferences. So if the concept of tomato explains how we categorize an object as tomato in typical cases, and we typically categorize tomatoes sometimes as a fruit and sometimes as a vegetable, the notion of concept does not seem to fulfill its role. Hence, we need more than one concept for the same category. 'Tomato' could thus be explained by two different stable concepts – one for each classification or contradictory judgement.

This strategy of cutting concepts into smaller subconcepts can be applied to other cases that

would normally support Contextualism. In every typical situation in which a subject seems to retrieve different features depending on the context, the Invariantist can posit not one context-depend concept, but two that are context-independent. This notion can be applied to the empirical example above. Instead of arguing that the concept of piano is flexible depending on the context, we can say that we have two piano concepts, one that determines our inference that pianos are musical instruments, and another that explains why, in another context, we infer that they are heavy.

This line of defense clearly adds to the difficulty to falsify Invariantism, but it may have another problematic consequence: according to Machery, it leads to Eliminativism. Machery (2009) argues that there is much evidence in the psychological literature that suggests what we commonly call 'concept' can be divided into at least three different kinds of fundamental bodies of knowledge, namely Prototypes, Exemplars and Theories. Combined with the argument that concepts are too coarse-grained for at least a subset of our categorization behaviors, he concludes that the notion of concept does not fulfill the explanatory role we ask it to play. More crucially, it may not pick out anything that actually exists in nature. Hence, the notion of concept should be eliminated from our terminology in psychology.

This conclusion deserves to be dealt with more carefully, but due to a lack of space I remain brief. For now, it will be enough to say that if Machery is right and Invariantism really leads to Eliminativism, which is a highly controversial view with many theoretical disadvantages (see the commentary on Machery, 2010a, 2009), Invariantism may be scientifically more disadvantageous than is commonly assumed. If Machery is right, it would not only eliminate an important tool in psychology, but could also have a similarly consequential effect for other fields, such as linguistics, philosophy and neuroscience that all take the notion for granted. Contextualism, on the other hand, does not require splitting up concepts into sub-concepts in order to account for context-effects, and does therefore not have this problematic consequence. It would thus not only be more parsimonious by proposing one instead of several concepts, but would also, perhaps surprisingly, be more conservative than Invariantism.

Although I cannot fully discuss this complex debate on Machery's Eliminativism, it adds to a growing suspicion that Invariantism does not seem as conservative and empirically supported as it is often presented. This leaves us with a difficult situation. If I am right, no empirical evidence can decide between Contextualism and Invariantism and at least in terms of explaining

stability, both views are equally successful. Still, we hope that somehow it will be possible to assess whether we should be Contextualists or Invariantists considering the important consequences of this decision for our theories of language and cognition. In the next two sections I defend a possible solution. I argue that Invariantism cannot account for the other central desiderata for any theory of concepts – scope and compositionality.

3. Scope

The *scope-desideratum* is a meta-theoretical criterion, according to which, everything else being equal, a theory with a wide scope has an advantage over a theory with a small scope (Prinz, 2002; Machery, 2006). Like other meta-theoretical considerations as parsimony, conservatism or productivity, the scope desideratum has an important epistemological function: a theory that has a limited scope is thought to fail to capture actual categories in the world. In the words of Thomas Aquinas, "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices" (1945, pp. 129). Applied to the present case, a theory that can explain similar phenomena in terms of one mechanism may have a better chance of "carving nature at its joints" than a theory that requires two.

I argue that Invariantists do not meet this criterion because they have difficulties accounting for superordinate and many genuinely abstract concepts.[13] This is highly problematic. According to the common notion of abstract concepts as concepts whose referents are physical objects that can be perceived by our senses, comparably few concepts are concrete (Kiefer and Pulvermüller, 2012, Wilson, 1988; Spreen and Schulz, 1966). According to a second popular definition that focuses on imagery, even fewer concepts would be considered concrete (Paivio et al., 1968). Moreover, it is widely agreed that abstract concepts comprise the majority of the vocabulary of adult languages (O'Grady, 2005; Gentner, 1982; Brown, 1957). Thus, a theory that can only explain concrete concepts would radically restrict its scope and its ability to explain central phenomena of higher cognition.[14]

13 According to the definition of abstract concepts as lacking physical references many superordinate categories should be considered concrete. According to other definitions, based on imagery for instance, they should not. Nothing hinges on this here, but I will adhere in the following to the latter understanding and consider superordinate categories abstract.

14 Note that even though many abstract concepts are highly complex, they still require an explanation in terms of a wide-scope theory of concepts. Unlike ad hoc categories that may be repudiated as not being the kind of categories that we need a theory of concepts for (Machery 2009), abstract concepts are established categories that we constantly deal with in everyday life and are not intrinsically context-dependent as, for example,

Consider for example the superordinate categories 'animal' and 'vegetable' and the highly abstract categories 'love' and 'art'. Both kinds of categories share an important property: even typical instances of their extension have little in common that could be captured by a stable concept. Consequently, both are extremely difficult to acquire in terms of one or two encounters with such instances. A German Shepherd shares many properties with other dogs (shape, behavior, color) and may thus be a typical exemplar of the category 'dog'. However, it does not share as many relevant properties with other animals (such as snakes or birds). It may still be a typical animal (which depends on one's environment), but this typicality is not as informative for the superordinate category as it is for the basic-level category 'dog'.

The problem that arises for Invariantism is the following: Invariantism critically depends on the idea that a small context-insensitive set of features can determine all of our cognitive behavior associated with a category (in the case of Definitionism) or at least our typical behavior (in the case of Prototype Theory). However, in the case of many abstract categories, such a set is extremely difficult to find. That which most or all instances of these categories have in common is either not accessible to perception or uninformative. In other words, the set of common features is not able to explain our typical higher cognitive behavior.

For example, members of superordinate categories like 'animal' have, by definition, fewer features in common than basic-level categories. The features that they do have in common, say, 'has a color', 'has a shape' etc., are either too broad, i.e., also apply to typical members of other categories (hence are not informative), or are usually not accessible to perception ('has a heart' or 'breathes' for example). In both cases, they cannot explain everyday categorization and induction behaviors, which require informativeness and perceptual access.

In addition, we do not always employ representations of the same typical features for different conceptual tasks. In the case of basic-level categories we may recognize a dog by its superficial features and also retrieve this information when hearing that Fido is a dog (say, it's shape). However, in the case of superordinate categories, the information we use in inferences and the information we use to recognize objects in our environment often differ. For example, we may use superficial features to recognize a potato as a vegetable, but do not usually immediately

indexicals (like 'I' or 'here') are. The meaning of the word art does not depend on context in the same way as 'I' or 'here' depend on context.

infer that vegetables look like potatoes (although some do).

To illustrate the extent of this problem, recall how Invariantists would account for concrete basic level concepts as DOG. For Invariantism to be true, the information that overlaps in all typical scenarios and inferences associated with dogs must be able to explain and determine at least typical DOG behavior in typical contexts. In other words, only that which typically overlaps and that which determines typical conceptual dog-behavior is to be called 'concept'. Everything else is, according to Machery, part of background knowledge. Applying this strategy to a superordinate category such as 'animal' turns out to be extremely difficult.

Imagine all the different properties that we may use to identify an entity as an animal, or that we infer if somebody uses the word 'animal' in a conversation. That which seems most informative about animals, perhaps that they all breath, walk by themselves, eat and sleep, is neither what we typically infer when somebody talks about animals, nor what commonly allows us to distinguish animals from other kinds. We seldom experience animals as breathing or eating, nor do we only recognize animals as such if they are walking or breathing. A dead animal is recognized immediately as an animal without doing any of these things. It also does not seem likely that we have acquired the concept of animal by inferring that the animals we had previously seen usually move by themselves and eat. It is, for example, difficult to imagine how children without exposure to real-life non-human animals would learn this category if it was reducible to these features.

Vegetable is another good example. The informative and highly typical properties that most vegetables share are 'plant', 'eaten with savory food' and 'healthy'. However, first, none of these categories seem to be any less problematic than 'vegetable'. It is not clear, for example, how we represent the notions of plant and healthy if vegetables are themselves part of both classes (the danger of circularity). Second, we usually do not recognize vegetables as vegetables by properties as 'healthy' or 'plant'. We typically recognize vegetables by their shape, color and other features accessible to perception. However, onions and potatoes share very few superficial features, so how can perceptually accessible common features explain that we recognize them both immediately as vegetables?

Similarly, 'love' is a category that is extremely difficult to understand in terms of a few overlapping properties. Love can be understood in terms of having both good and bad feelings,

as well as having healthy and destructive interactions with another person. Love is associated both with the feelings of trust and insecurity and many other contradictory properties. In addition, the set of behaviors that we may associate with the concept of love is extremely diverse. We may identify an instance of love if a couple is very close, but we may also say that an elderly couple loves each other even if they hardly talk or show signs of love that we would usually associate with younger couples.

Finally, if Invariantism were true, the category of art would be one of the most mysterious. There is very little that even typical artworks have in common and, again, the properties based on which we identify an artwork may not be the same as the ones we typically infer when hearing about art. For example, a typical sculpture and a typical painting may have little in common. What they could have in common is that they might be exhibited in an art gallery. This can hardly be used to reliably pick out artworks even in typical cases, since the property of being in an art gallery is shared with many objects that are not considered art, such as radiators, people and chairs.

There is also empirical evidence that supports the idea that superordinate categories are not primarily represented by what its instances have in common, such as 'being healthy' in the case of vegetable, 'breathes' in the case of animals or 'being exhibited in art galleries' in the case of art. Findings by Lebois et al. (2015), Blanchette & Dunbar (2000) and Forbus, et al. (1995) suggest that peripheral information of a category is activated before less superficial features are activated. Essentialist information (e.g., that all animals breath), the studies suggest, may thus only be activated when context demands it. This shows that either automaticity is not a positive outcome of context-independence or it is not evidence for it. Both claims however, are essential to Invariantism (Machery, 2015, 2009).

Contextualists, on the other hand, have a relatively straightforward answer to these problems. Since they deny that concepts are reducible to those features that overlap in all or typical situations, they can allow that features that do not overlap can explain typical higher cognitive behavior and hence be concepts. For example, instead of requiring that the concept of vegetable is a complex of representations of what all or all typical vegetables have in common, Contextualists can argue that we use different features in different situations to make certain inferences and categorizations. Although typical instances may be recognized faster, this does not mean that the notion of concept is to be reduced to such features. What makes all these

different representations the concept of vegetable is, to name just one option, that they reliably correlate with vegetables (Prinz, 2002; Barsalou, 1999; Fodor, 1998; Rey, 1983).

However, none of this means that Invariantism is not at all equipped to explain abstract concepts. Machery and others could respond in two ways. First, they could give up a key assumption of their view. They could, for instance, argue that we need background knowledge even in typical contexts to adjust our stable context-independent concepts to the context. This would mean that even behaviors like identifying a potato as a vegetable or saying that a dog is an animal would require background knowledge. This is an extremely unusual way to speak about concepts. We would expect that at least in typical cases background knowledge is not necessary.

A second option is to cut concepts into smaller pieces, i.e., to posit sub-concepts – one for each context or class of typical contexts, as Machery (2009) proposes. In the case of 'tomato', Machery claims that we do not have one, but two concepts of tomato; one that allows us to identify tomatoes as a vegetable and the other to classify them as fruit. For the category of vegetable, this would mean that we had to posit a large number of concepts of vegetable, one for each typical context in which we need a different set of features. In the case of LOVE, we might have a concept of love for the young couple, for the older couple, for unrequited love, and so forth.

This option runs into the problem that we do not have the impression that we are really talking of different concepts of love when talking about young as opposed to older couples. Rather, it seems more natural to think that we use different typical properties or features of this category to apply the concept of love in different situations. Similarly, we do not commonly have the impression that we are talking about different concepts when thinking about animals in the context of the desert and in the context of a big city. Instead, we seem to have the same category in mind only that it is more natural to think first of dogs rather than snakes if we live in big cities, although both are equally good exemplars of this category.

A further problem would be that an important distinction between homonyms (an expression that is attached to different concepts) and expressions that have only one meaning that can be used in very different situations would evaporate. An expression like 'art' or 'animal' would be associated with many different concepts just as 'bank' is associated with the concepts of

riverbank and financial institute. However, this seems highly implausible as, at least intuitively, there does seem to be an important difference between words like 'animal' and homonyms like 'bank'.

Finally, and most importantly, both of these options fail because they would make Invariantism collapse into Contextualism. In the case of the first option (allowing background knowledge to explain even typical situations), this is because even in typical situations, context-dependent representations would be required to explain typical higher cognitive behavior. This is precisely what Contextualists claim and what Invariantists essentially reject.

In the case of the second option (positing several concepts per category), Invariantists would have to propose an often large set of sub-concepts for the same category (think of all the different kinds of typical instances of art, love, animal and vegetable). So what would eventually explain the behavior associated with the same category are the sub-concepts that are retrieved depending on the demands of the respective situation or context and not the concept representing the category. Again, context would be required to explain typical behavior, which is identical to the proposal of Contextualists.

4 Compositionality

The same difficulty arises for a second important constraint on any theory of concepts – c*ompositionality*. A theory of concepts has to explain how they can combine in a logical and systematic way in order to explain how thought can be productive, i.e., how we can form an unlimited number of distinct thoughts with a limited number of concepts (Prinz, 2002; Fodor, 1998; Evans, 1982).

Compositionality is generally raised as a serious challenge to psychological theories of concepts. The classic example to illustrate this problem is the concept of pet fish. Typical pet fish cannot be said to possess any of the typical features of its constituents. PET (e.g., ‘is furry’, ‘lives in the house’, ‘can be cuddled’) and FISH (e.g., ‘lives in the ocean’, ‘has scales’, ‘is large’) combined do not produce the prototype of pet fish, which have completely different characteristics, such as 'lives in a tank' or 'is small and colorful'.

Another famous example is WOODEN SPOON. Stable representations have difficulties explaining this concept, because when subjects are asked what a wooden spoon is they do not

simply replace ‘metal’ with typical features of ‘wooden’. Instead, they automatically change other features of the concept of spoon, such as 'slightly bigger than spoons' or 'used for frying pans' (Kamp and Partee, 1995).

These examples however cannot challenge Invariantist views, such as Prototype Theory, per se, since both PET FISH and WOODEN SPOON are established concepts that we are already familiar with. In a sense, they presuppose what needs to be shown (Prinz, 2012). From the observation that wooden spoons are usually not understood in terms of the conjunction of 'spoon' and 'wooden', it does not follow that Prototype Theory is false. Both concepts simply refer to entities that are different from what one would expect from the common use of the words 'wooden' and 'spoon'.[15]

Since an explanation of how concepts compose is primarily required for an explanation of the productivity of thought, the important question is not how we explain established combinations, but how we combine concepts to generate combinations that we have never encountered before (Machery and Ledere, 2012). It would thus be more surprising if Prototype Theory did predict what wooden spoons actually look like since this knowledge is not available to us if we do not yet know what is meant by the expression. It is more reasonable that WOODEN SPOON is a new concept that may have its origin in the concepts of wooden and spoon, but that is not reducible to them.

If it is crucial to show how concepts combine when the combination is not yet known to us, an Invariantist view that is based on typicality, for example, would have an advantage that is often not acknowledged by opponents of psychological theories of concepts (Fodor, 1998, for example). If we do not know anything about the intended combined concept, we generally start out by combing typical, not atypical features. Unfortunately, this initial advantage for Prototype Theory turns into a serious disadvantage for Invariantism. The problem is that even if both constituents of the combined concept are represented by one stable set of features per concept, many combinations of complex concepts are logically possible.

Take for instance again the concept of pet fish. The advantage of Prototype Theory was that it can explain how we make combinations that yield completely new ideas out of old ones. But

15 Both examples could still be effective if Prototype Theory was understood as a kind of Conceptual Role Semantics (see Fodor, 1998).

how do we combine PET FISH if we had never heard of this combination before? If the concept of fish includes information as 'lives in the ocean', 'swims', 'has gills' and if PET is typically constituted of ‘furry’, ‘cuddly’ and ‘domesticated’, which attributes should we combine? 'Pet fish' could mean 'furry domesticated gill breather that lives in our garden' or 'cuddly dog-shaped animal that swims in the ocean'. This complexity intensifies if we acknowledge that there are more than one or two typical pets and fish. In this case the variety of possible combinations increases dramatically.[16]

To explain this phenomenon, Invariantists have again two options. First, they could argue that we retrieve every logically possible combination as an individual concept. In this case, we would require a context-dependent selection of the appropriate combination and context would be necessary to explain even typical situations. A second option is that we combine only some features and then allow additional features to be selected from background knowledge to adjust the concept to the context. Both options require a selection of features in a context-sensitive way. As before, Invariantism would collapse into Contextualism.

In the case that Invariantists opt for the first option, according to which we construct all possible combinations by default, i.e., independent of context, we would still need to select the combination that makes the most sense or is the most appropriate in the respective situation. Since appropriateness is determined by context, the explanation of the behavior even in typical cases would include context and Invariantism would lose its essential characteristic. For example, if we wanted to combine PET and FISH, the first option predicts that we combine all the logically possible combinations of the constituents of PET and FISH. This will generate a large number of different concepts. For instance, if the stable concept of each category includes three default features, it would generate 27 different concepts. Only context can determine which is the most appropriate.

In the case that Invarianists opt for the second option, i.e., to model the respective behavior by positing the construction of one or two concepts (perhaps retrieved at random from the possible combinations), we would need background knowledge to adjust this retrieved knowledge to the context unless, by chance, we would have generated the correct concept. This chance, however,

16 Note that combined concepts should not be confused with Barsalou's (1987) ad hoc concepts, which are intrinsically context-dependent and thus repudiated by Machery (2009) for not really being the kind of concept a theory of concepts needs to explain. Combined concepts are not intrinsically context-dependent. Context could theoretically play a role only after the combinatorial process is completed.

is small if each concept contains three features, namely 1 in 27 (3.07%) and can thus only explain few cases. So again, Invariantists would need context-dependent background knowledge to make the adjustment and explain the observed behavior. They would thus have to justify why context-dependent knowledge should be excluded from the notion of concept.

Interestingly, the latter strategy is usually preferred by many leading Invariantists. A recent paper on this issue by Machery and Lederer (2012) provides a good overview of the possible positions, none of which denies the importance of context. It also presents a detailed response to the problem of compositionality, which is based on several heuristics or "rough and ready" strategies for producing the appropriate combination of concepts. These heuristics all put a strong emphasis on context and active construction of a concept, as opposed to passive retrieval as Invariantists predict.

For example, according to Machery and Lederer's (2012) *Modality-Specific Heuristic*, a complex phrase "is almost always embedded in a discourse or narrative context that specifies its intended modality—that is, the discourse or narrative context specifies the modality, or perceptual sense, to which its meaning is relevant" (pp. 75). This means that the subject only chooses those features for a combination that are related to the contextually appropriate modality. For instance, if the context involved a painting, the combination HALF GRAPEFRUIT will retrieve the feature 'pink' but not 'sour', as the taste is not required in the context of a painting. This explanation of compositionality can justly be called Contextualist.

What does this mean for Invariantism? It means that generalizations of behavior in which compositionality is a requirement demands context-depend selection of either the right concept or the context-dependent selection of the right features to be combined. Since compositionality is one of the key desiderata for any theory of concepts and since in order to account for compositionality Invariantists have to include context, Invariantism again collapses into Contextualism or fails to account for the large class of composed concepts.

5 Conclusion

Concepts in cognitive science are posited to explain predictable higher cognitive behavior, such as categorization and inference-making, in a complex and changing world. It is thought that concepts explain this behavior because they mirror the regularities and commonalities among the entities we encounter in our environment. Both Contextualists and Invariantists assume that

some information associated with a category may overlap independent of context. Both also agree that we need additional information to explain appropriate behavior in unusual contexts. The disagreement is about whether the information that overlaps can explain our typical and regular behavior or not. Invariantists are committed to this idea and Contextualists reject it.

Whether we should be Invariantists or Contextualists thus seems to be ultimately an empirical question. However, I argued that the empirical evidence that appears to support Invariantism, namely certain frequency and typicality effects, can easily be explained by Contextualists. On the other hand, evidence that could support Contextualism, namely irregularity and context effects can be explained by Invariantists by either emphasizing the role of background knowledge or by splitting up seemingly context-sensitive concepts into several context-independent sub-concepts.

Instead of focusing on empirical evidence, I suggested returning to theoretical arguments. In particular, I argued that Invariantists cannot account for two important desiderata of any theory of concepts: scope and compositionality. In the case of abstract concepts, because that which is typical for these categories cannot explain our typical higher cognitive behavior. In the case of composed concepts, the possibilities of combination of even typical features are so large that context is needed to disambiguate the notion.

This leads to a dilemma. Invariantists could only account for abstract and composed concepts by either referring to background knowledge or splitting up context-sensitive concepts into several sub-concepts. In both cases, Invariantism would be indistinguishable from Contextualism. Moreover, Contextualism seems to have several methodological advantages. First, it retains the distinction between homonyms and complex concepts. Second, it avoids the theoretically problematic consequence of at least Machery's version of Invariantism, i.e., Eliminativism.

Chapter 3: Embodied Cognition and Abstract Concepts: Do Concept Empiricists Leave Anything Out?

Abstract: According to the embodied cognition hypothesis, the mental symbols used for higher cognitive reasoning, such as the making of deductive and inductive inferences, both originate and reside in our sensory-motor-introspective and emotional systems. The main objection to this view is that it cannot explain concepts that are, by definition, detached from perception and action, i.e., abstract concepts such as TRUTH or DEMOCRACY. This objection is usually merely taken for granted and has yet to be spelled out in detail. In this paper, I distinguish three different versions of this objection (one semantic and two epistemic versions). Once these distinctions are in place, we can begin to see the solutions offered in the literature in a new, more positive, light.

1 Introduction

According to the embodied cognition hypothesis (Barsalou, 1999; Pulvermüller, 2013; Prinz, 2002, 2005; Gallese & Lakoff, 2005) cognition is realized by the same cognitive and neural systems that are responsible for sensorimotor and emotional processing (from now on called 'concept empiricism'). This means that a subset of the perceptual, motor, emotional and introspective states that arise during encounters with, say, chairs will be stored in long-term memory to stand in as a symbol for chairs. Such *modal symbols* have both content and causal powers and can thus function as mental representations in the traditional sense (e.g., Pylyshyn, 1980). The main difference between modal and amodal symbols is that the former are analogous to the sensorimotor input that caused them and that their manipulation is more akin to a *re-enactment*, *simulation* or *emulation* rather than a passive retrieval (Machery, 2007).[17]

According to the competing view (e.g., Landauer & Dumais 1997; Machery, 2007; Mahon & Caramazza, 2008) conceptual representations are amodal and conventional. This means that sensorimotor input is transduced into a completely different mode of representation and is thus detached from the perceptual, motor and emotional systems that originally produced it. An example of such amodal conventional representations are linguistic symbols that, in most cases, bear only an arbitrary relation to their contents. Another example are conceptual mental representations, such as feature lists, frames or semantic networks, all of which store a kind of meta-data, e.g., in the inferior temporal or the prefrontal cortices, of what was originally produced in perceptual brain areas (Machery, 2016).

Due to the arbitrary relation to their contents, amodal symbols are considered to have their wide scope essentially. Like words, amodal symbols can, in principle, represent any kind of entity in the world, be it a concrete category like *table* or a more abstract category like *democracy*. However, the arbitrariness of amodal symbols raises the question of how these representations receive their content (what Harnad, 1990 calls *the symbol grounding problem*). One option is that they receive their content via mediating modal symbols (e.g., Harnad, 1987; Paivio, 1991; Margolis, 1998). Another option is that modal symbols are combined in such a way that they constitute more abstract ptamosal symbols, for example in so-called convergence zones (Damasio, 1989).

[17] I use capital letters to denote concepts, single quotation marks to denote words and italics for technical terms and to denote properties.

Concept empiricists have argued that the step of transducing meaningful sensorimotor symbols into amodal symbols is methodologically problematic (Barsalou, 1999). First, the nature of the transduction, i.e., both the encoding and decoding of the symbol, remains unclear. Secondly, by eliminating the need for transduction, concept empiricists claim to avoid the symbol grounding problem in a powerful and more parsimonious way. So, concept empiricists do not deny that amodal symbols can in principle account for the same phenomena as modal symbols, but rather that modal symbols have, first of all, a methodological advantage.

Opponents of the embodied cognition hypothesis have disagreed that modal symbols have a methodological advantage because, allegedly, they have difficulties explaining how we can represent ideas that we do not have sensorimotor access to. This would limit the scope of modal symbols and with it its supposed methodological advantage (thus, it has been called 'the scope objection'). How, for example, are we able to think about *the number four* or *truth* if our explanatory devices are symbols that are derived only from direct sensorimotor-emotional experience? So, it may be that concept empiricists can provide a relatively straightforward account of concrete concepts like CHAIR or DOG, but it is difficult to imagine how they could explain our ability to think about more abstract properties.

This objection has been extremely influential and even convinced many concept empiricists to opt for a pluralist view (Dove, 2009, 2016; Meteyard, et al., 2012). For example, according to Kiefer and Pulvermüller (2012, p. 820):

> By definition, abstract concepts do not refer to physical objects that can be directly experienced by the senses and their action relationship is, if it exists at all, very complex. At the first glance, it is therefore hard to imagine how such concepts could be grounded in the sensory and motor brain systems. Hence, the mere existence of abstract concepts appears to falsify modality-specific theories and points to an amodal symbolic representation.

Similarly, Mahon and Caramazza (2008, p. 60):

> Concepts of concrete objects (e.g., HAMMER) could plausibly include, in a constitutive way, sensory and motor information. But consider concepts such as JUSTICE, ENTROPY, BEAUTY or PATIENCE. For abstract concepts there is no

> sensory or motor information that could correspond in any reliable or direct way to their 'meaning'. The possible scope of the embodied cognition framework is thus sharply limited up front; at best, it is a partial theory of concepts since it would be silent about the great majority of the concepts that we have.

It is worth emphasizing how potentially devastating the objection from abstract concepts is. Abstract concepts in psychology and psycholinguistics are usually defined in two ways (Hoffman, 2015), both of which render abstract concepts extremely common. First, abstract concepts are defined in terms of the psycholinguistic variable *abstractedness*.[18] Briefly, abstractedness is a measure of the degree to which a concept picks out entities that one can touch, see, hear or smell (Spreen and Schulz, 1966; Borghi and Cimatti, 2009). DEMOCRACY is usually rated as highly abstract, in this sense, while TABLE usually receives a lower rating. Secondly, abstract concepts are defined in terms of another psycholinguistic variable called *imagery* (e.g., Vigliocco et al., 2014, Paivio, 1991). Imagery is operationalized in terms of how well subjects can retrieve images of instances of the respective category.[19]

According to both definitions, abstract concepts are extremely common. Consider how few concepts denote physical objects that can be perceived by our senses. In particular, concepts that are essential to our social lives would lie outside the scope of concept empiricism (JUSTICE, LOVE etc.). Measured in terms of imagery, even fewer concepts are to be considered concrete. For example, since superordinate categories ('animal', 'vegetable', 'thing') cannot, by definition, be represented in terms of an image (Lakoff, 1987), they too would fall outside the scope of perceptual symbols. In other words, if concept empiricists cannot explain abstract concepts, they could only account for beliefs about existing concrete middle-sized everyday objects located on the basic and subordinate level.

Because abstract concepts are considered extremely problematic for concept empiricism, many concept-empiricists have begun to opt for weaker, pluralist versions of concept empiricism (Pulvermüller, 2013; Vigliocco, et al., 2004; Dove, 2009; Meteyard et al., 2012). According to such *symbol pluralism*, either concept empiricists can only explain concrete concepts or most concepts require both modal and amodal symbols. By embracing pluralism we could explain

[18] What Dove (2016) calls *dis-embodiment*. Abstractedness lies on the same continuum as *concreteness*.
[19] Both ways of defining the notion of abstract concept in psychology are not ideal, but since it is the way the notion is defined in the relevant scientific literatures I stick to this use for this paper's purpose. Note that nothing hinges on the question of whether this definition is adequate here.

embodiment effects for concrete concepts and still employ amodal symbols to explain abstract concepts.

There are, however, strong reasons to resist such an early pluralism. First, consider again that one of the selling points of modal symbols was that they are supposedly methodologically advantageous (Barsalou, 1999, p. 580). To posit two kinds of symbols would eliminate this initial methodological advantage. Secondly, pluralism inherits the problems of amodal symbols that were mentioned above, e.g., that the process of transduction remains unclear. Thirdly, wide-scope concept empiricism is still productive. Although research on abstract concepts is still in its infancy, there is already some promising behavioral and neuroscientific (fMRI, TMS) evidence that suggests embodiment effects (intermodal transfer costs, modal facilitation, modal interference) for social categories like 'convince' (Wilson-Mendenhall et al., 2013), emotion words (Vigliocco, 2014), action words (Pulvermüller, et al., 2005) and even numerical concepts (Bergen, 2012; Lindemann & Fischer, 2015).

Although none of this preliminary evidence is decisive, and although there have been some objections to the empirical evidence put forward by concept empiricists (Machery 2007), the existing embodiment effects for abstract concepts should suffice as a motivation to refrain from ruling out the empirical possibility of a wide-scope concept empiricism. Furthermore, it should be noted that the vast majority of empirical work on this subject is still focused on the relatively small set of concrete concepts (mostly concrete nouns and some verbs). With more research on abstract concepts we may get a better picture of how abstract concepts are represented.

The aim of this paper is to make room for the possibility of a wide-scope concept empiricism by defending it against the scope objection. After introducing the scope objection in more detail, I distinguish two issues that ought to be kept apart when theorizing about concepts, i.e., issues pertaining to content determination and issues pertaining to concept application. I argue that most of the objections to concept empiricism address the first issue, while psychologists are more interested in the latter issue. Finally, I argue that most of the philosophical objections to wide-scope concept empiricism presuppose a descriptivist theory of content, which has been challenged by numerous philosophers of language (e.g.., Putnam, 1975).

Put differently, contemporary concept empiricists (like Barsalou, Pulvermüller and Vigliocco), unlike traditional empiricists (Locke, Hume) or logical empiricists (Carnap, Schlick) are usually

not concerned with the view that all knowledge is derived from experience (but see Prinz, 2002). While traditional empiricists are primarily concerned with questions of content, i.e., descriptive and foundational semantic questions (i.e., questions of what our symbols mean and how they get their content), contemporary concept empiricists in cognitive science presuppose such a theory of content (Barsalou, 1999, for example, presupposes a causal-historical theory of content) and are primarily concerned with the question of how we apply and acquire everyday concepts.

2 The Scope Objection

To save concept empiricism, Barsalou, Prinz and others have tried to show that abstract concepts can, at least in principle, be represented in sensorimotor-introspective and emotional systems. For example, Barsalou (1999) and Prinz (2002) proposed that even a highly abstract concept like TRUTH can be explained in terms of introspecting a mental operation that compares beliefs with currently perceived events. A "match" would be conceptualized as a true statement. DEMOCRACY, Barsalou and Prinz argue, can be explained in terms of concrete situations of voting combined with the feeling of freedom. Another way abstract concepts have been explained is in terms of metaphors. Lakoff and Johnson (1980), for example, claim that our concept of argument is shaped by the metaphor *argument is war*. This is supposedly reflected in our use of expressions such as "one defends a position" or "one attacks a claim". They propose that this is linguistic evidence for the idea that we represent the former concept in terms of the latter metaphor and thereby reduce its degree of abstractedness.

Opponents of the embodied cognition hypothesis have not been convinced by these proposals and even some of its proponents have become skeptical that a wide-scope concept empiricism is possible, opting for weaker versions of embodied or empiricist theories of concepts (for a review see Meteyard, et al., 2012). Mitchell & Clement (1999), for example, argue that Barsalou's analysis of TRUTH fails because a match between a visual simulation and a visual experience could represent not only *truth*, but also *similar, comparable* and *looks like*. Similarly, Adams & Campbell (1999) argue that we could imagine cases of *matching* that have nothing to do with *truth* (when playing Tetris, for example). In other words, both argue that the process of matching is neither necessary nor sufficient to capture the meaning of 'truth'.

Siebel (1999) points out that if Barsalou were right we would always apply the concept of truth whenever our beliefs match the contents of our perceptual states. This, he argues, would make

it impossible to understand illusions or hallucinations. Moreover, we would always detect falsity when perceiving a mismatch even though the respective proposition may in fact be true. In other words, accounting for TRUTH in terms of MATCHING does not explain many typical behaviors or capacities we associate with TRUTH, such as our capacities to detect falsity and illusions.

Ohlsson (1999) makes a slightly different point. He argues that even if we were able to represent or detect truth from "matches", we would not be able to detect falsity from a mismatch. This is because, according to Barsalou and Prinz, the proposition "the cat is on the mat" would be considered true only as long as my sense datum matches my belief that the cat is on the mat. However, if the cat is no longer on the mat, we cannot rely on the operation of matching to detect that the above proposition is false, as the absence of evidence is not evidence for absence (observing the sky and not seeing aliens does not mean that we are alone in the universe).

Dove (2009) raises a similar objection to an explanation of abstract concepts in terms of metaphors. He argues that metaphors highlight the similarities between two categories, say 'argument' and 'war', but not their differences. Dove reminds us that wars and arguments are obviously different in important respects, just as freedom is not always identical to the concept of lack of physical restraint. Moreover, Dove questions whether the existence of certain linguistic expressions like "winning an argument" is evidence at all for the metaphorical representation of our concepts because it is not clear whether such behavioral practices reflect conceptual and not just linguistic structures.

With respect to Barsalou's and Prinz' explanation of DEMOCRACY in terms of the event of voting and the feeling of freedom, Dove (2009) also objects that perceptual aspects of events of voting cannot sufficiently track democracies because they cannot distinguish genuine from non-genuine acts of voting. Moreover, he argues that even if we do not have sufficient knowledge to track democracies, we are still able to think about a country in terms of it being a democracy. For example, we can think that Moldova is a democracy without knowing anything about this country, especially how elections are held or how power is transferred.

At first glance, it may seem that all these objections raise roughly the same objection:
concept empiricists' attempts to account for abstract concepts are insufficient because they leave too much unexplained. Moreover, it seems that current explanations of abstract concepts

by concept empiricists have failed in such an obvious manner that it is rather implausible that future attempts will be more successful. The aim of the remainder of this paper is to show that this conclusion is based on a confusion between semantic questions of content and psychological questions of concept application or categorization and that even the simplistic interpretations of concept empiricists' explanations of abstract concepts cannot be refuted by the above objections.

3 Content and Concept Application

Concepts are the constituents of propositional attitudes, such as beliefs and desires (i.e., thoughts). This is relatively uncontroversial (Fodor, 1998; Margolis & Laurence, 2014) and even assumed by many philosophers who defend the ontological view that concepts are abilities (e.g., Liptow, 2012). One reason to think that thoughts are structured is that they are productive and systematic (Fodor, 1989), which means that we can form an unlimited number of new thoughts despite having limited mental resources. The hope is that we can explain this phenomenon if we assume that propositional attitudes have constituents (concepts) that can be combined in unlimited and systematic ways.

Concepts are at least partly individuated by their *semantic content* (roughly what Frege (1892/1984) called *Bedeutung* or *reference*). We know that DEMOCRACY and TABLE are different concepts because they are about different things, namely *being a democracy* and *being a table*. Concepts may also need to be distinguished by means of their *epistemic content*, i.e., roughly, that which people believe about democracies and tables (which Frege called *Sinn*).[20] For instance, the concepts MORNING STAR and EVENING STAR may have the same semantic content (reference) but different epistemic contents (senses).

Questions about the meaning or content of a concept (so-called 'descriptive semantic issues') need to be distinguished from so-called 'foundational semantic issues' of content (Lewis, 1970; Stalnaker, 1997), which attempt to answer the more metaphysical question of what makes it the case that concepts can be about one thing rather than another. In the concept literature, philosophers generally distinguish two such foundational theories of content. According to semantic internalists like Peacocke (1992), that which determines the content of a given concept

[20] Frege also used the term *mode of determination* (*Art des Gegebenseins*), which is supposed to be mind-independent. However, in many contemporary philosophies of concepts, Frege's notion of sense is mentalized. Nothing here hinges on this distinction.

is its relation to other concepts. For example, that which determines the content of the concept of bachelor is its relation to UNMARRIED and MAN. According to semantic externalists (e.g., causal-historical theories), on the other hand, content is determined not by means of descriptions (or beliefs), but by external (e.g., causal) relations between one's mind and the respective object, property or relation (Kripke, 1972; Putnam, 1975; Millikan, 1998; 2017; Fodor, 1998).

Foundational issues of content are important for theories of concepts not only because they explain how a symbol can have semantic content, but also because they determine a concept's possession conditions. According to semantic internalists, I possess the concept of bachelor if and only if I am able to draw the inferences that determine the content of BACHELOR, e.g., if I know that if someone is a bachelor he is not married. According to semantic externalists, I do not need to have many true beliefs about bachelors to be able to have propositional attitudes about them. Instead, what is necessary and sufficient for being able to think about someone as having a certain property or relation is to be causally connected to this property or relation in an appropriate way. According to semantic causal-historical views, this referential causal relation can be established externally, e.g., by a simple causal perceptual mechanism or by relying on experts in ones' linguistic community (Burge, 1979; Margolis, 1998; Margolis & Laurence, 2011).

Semantic accounts of concepts (both descriptive and foundational) need to be distinguished from accounts of how concepts are applied in categorization and other higher cognitive behavior, which are primarily investigated by psychologists (Machery, 2009). Psychologists usually presuppose one or both of the above semantic accounts and instead focus on what they think are psychologically more relevant questions, such as how we learn and use the *common* (as opposed to the *correct*) application conditions of a concept that are relevant in the relevant linguistic community (Rey, 1983; Fodor, 1998; Machery, 2009). That which enable us to apply concepts can be called a *categorization device* (see Löhr, forthcoming).

Our categorization devices, i.e., that which allows us to make categorizations, are individuated not by their epistemic and semantic content but by means of their explanatory power (Machery, 2009). A categorization device of dog for instance could be a set of beliefs that we use to apply the concept of dog to instances even if this set of beliefs fails to pick out dogs and only dogs,

i.e., the referent of DOG.[21] The semantic content of a concept (its extension or referent) is thus to be distinguished from its epistemic content (e.g, certain beliefs that determine the referent or explain Frege Cases), and also from the mechanisms we use to apply concepts, i.e., that which constitutes our categorization devices.[22]

Accounts of concept application are usually proposed and empirically tested by psychologists and psychology-minded philosophers like Joshua Knobe (2013), Jessy Prinz (2002) or Edouard Machery (2009). Influential psychological theories are, for instance, *prototype theory*, *theory theory*, *exemplar theories* or *frame theory* (for reviews see Murphy, 2002 or Machery, 2009). The basic idea that these theories have in common is that cognitive systems classify their environment by storing certain descriptions, features, theories, images or sets of beliefs. Acquiring a categorization device (as opposed to acquiring a concept) just means learning what these descriptions/theories/exemplars are.

This *categorization internalism* is defended by all contemporary psychological theories of concepts and is highly intuitive. Except on a very low level of categorization (e.g., the early visual system), we expect that we recognize things in the world as belonging to a certain category because we hold beliefs about the criteria for belonging to the respective category. For example, it is highly plausible that we recognize trees as trees because we think that trees typically have a certain shape and perhaps also certain less superficial features investigated by biologists. These beliefs allow us to make certain inferences, e.g., that when something is a tree it is a plant and that it probably has leafs in the summer.

Categorization internalism is often conflated with *semantic internalism*. In particular, it is often argued against categorization internalism (like prototype theory, e.g., Rosch, 1975) that they violate necessary conditions for a theory of concepts. Most famously Fodor (1998), but also more recently Rice (2013), object that prototypes cannot compose productively, thereby failing to meet the standards for counting as a theory of concepts. For example, according to Fodor, the typical features of FISH and the typical features of PET do not produce the typical features of the concept PET FISH. Another objection to prototype theory by Fodor is that information

[21] Note that, a categorization device could, in principle, be composed of other mechanisms besides beliefs as long as this explains how we in fact classify our environment

[22] Note that some authors (e.g., Prinz, 2002) give a unified account of all three notions. However, this is not necessarily advantageous as I show in the following.

that can be stored in a prototype is too unstable to explain how people can share concepts (but see e.g. Prinz, 2002; 2012; Prinz & Clark, 2004 for replies to this challenge).

While it is more or less widely agreed that we should expect an account of concepts to explain how they can compose and be shared, it is not clear why we should demand the same from our categorization devices. On the contrary, because the individuation criteria of concepts are fundamentally different from our criteria for individuating categorization devices (Löhr, forthcoming), the satisfaction conditions of a theory for each are different. Again, concepts are individuated by means of their content, while categorization devices are individuated by means of whatever best explains our higher cognitive abilities (categorization, induction etc.). Consequently, while the former is meant to explain how thought can be productive, systematic and public, the latter ought to describe and explain how we *actually* combine our concepts and beliefs in order to adapt to new circumstances in our environment (by learning certain heuristics for instance, see Machery and Lederer, 2012).

How to individuate concepts is thus ultimately a conceptual question. What the common application conditions of concepts are, however, and how we actually apply concepts (i.e., what constitutes our categorization devices) is ultimately an empirical question to be investigated by psychologists. In other words, which descriptions we use in order to conceptualize objects can and usually does depend on the individual as long as their epistemic and especially their semantic content remain the same. Consequently, while, at least according to many philosophers including Fodor, concepts need to compose and be shared in order to satisfy our demands on a theory of the productivity of thought, that which allows us to apply concepts (perhaps prototypes or simulations of past experiences) need not.[23]

That we should draw a distinction between accounts of concepts and categorization devices can further be supported by Kripke/Putnam style thought experiments. The main argument against semantic internalism was that at least in some cases we seem to be able refer to or think about an entity without having many true beliefs to distinguish it from members of other categories. For example, many philosophers and psychologists have found it convincing that we can refer to Aristotle even if all or most of our beliefs about him are false (*semantic problem of error*) or incomplete (*semantic problem of ignorance*). If everything I associate with the name Aristotle

[23] E.g., a blind person uses very different descriptions to recognize something as a tree than a sighted person.

is that he was the teacher of Alexander the Great and this turns out to be false, as a historical fact, many share the intuition that I have said something false about Aristotle and not something trivial about who actually taught Alexander the Great.

We can apply the same arguments to illustrate the distinction between concepts and categorization devices even if we disagree with them on a semantic level (i.e., even if we disagree with Kripke and Putnam). So even semantic internalists (like Jackson, 1998 or Peacocke, 1992) can agree that in order to refer to a class of objects, we need not necessarily be able to identify instances of this class even in typical circumstances. This would be the case, for example, if we only knew certain essential characteristics of water but not its superficial ones (what I call 'the categorization problem of error'). Similarly, we can also be said to know that water is essentially H2O, but still form the false belief that water does not make up most of the substance we call 'coffee'[24] (what I call the 'categorization problem of ignorance'). Consequently, a semantic account of concepts does not automatically provide an account of our categorization devices, just as our categorization devices need not necessarily account for the semantic or epistemic content of our concepts.

Finally, because concepts and categorization devices have different individuation criteria, they also have different possession conditions. While, again, the possession conditions of concepts derive from our preferred fundamental theory of content, the possession conditions of our categorization devices depend on our individuation criteria of our categorization devices. Unfortunately, there is next to no discussion on the possession conditions of categorization devices, even though there is of course a very large debate on the possession conditions of concepts. To my knowledge, the only explicit formulation of individuation conditions of categorization devices (which determine its possession conditions) can be found in Machery (2009, 2010), although an alternative account can be derived from the literature on what I call *categorization contextualism* (Barsalou, 1999; Prinz, 2002; Löhr, 2017).

According to Machery, categorization devices are individuated in the following way: two sets of beliefs about the same class of objects are distinct categorization devices if a) they lead to contradicting inferences or b) if one set of beliefs is not immediately retrieved when the other is retrieved. For example, according to Machery, many people have two fundamentally different categorization devices of *tomato* because they may in one context immediately retrieve the

[24] See Malt (1994).

belief that tomatoes are vegetables, and also, in a different context, that they are fruit. So, according to Machery, since categorization devices are defined as that which explains the application of concepts and since we often have contradictory beliefs about the same category or kind (e.g., about 'tomato'), we possess at least two categorization devices of this category.

It is relatively easy to turn these individuation conditions of categorization devices into possession conditions. It seems that, according to Machery, in order to possess at least the typical categorization device of tomato (e.g., the prototype of tomato), I need to be able to immediately classify a tomato as, say, a red vegetable (even if, strictly speaking, this belief is false). I possess the other typical categorization device of tomato if I also readily classify tomatoes as fruit in other contexts. According to Machery, in this case I have then two stable categorization devices of the same concept as opposed to, e.g., one context-dependent categorization device (as argued by categorization contextualists).

Following Machery's possession conditions, we can see that possessing one or both of the common categorization devices of tomato and possessing the concept of tomato have very different requirements. The most important difference for now is that in order to possess the latter I do not necessarily require the former. Similarly, in order to possess the former, I do not necessarily need to possess the latter (although I do need certain other concepts such as RED and VEGETABLE). In other words, I can classify red vegetables as something tomato-like and even call it 'tomato' and still not meet the requirements for concept possession (depending on one's foundational semantic theory).

In the remainder of this paper I apply the distinction between semantic accounts of concepts and accounts pertaining to categorization (i.e., concepts and categorization devices) to the debate on the scope objection to concept empiricism. The main advantage of the distinction is that we can now clearly distinguish between different versions of the scope objection, which allows us to separate the empirical from the conceptual problems of concept empiricism. This will especially aid those in the cognitive science community who are interested in embodied cognition and concept empiricism as empirical hypotheses (e.g., Barsalou, Pulvermüller, or Vigliocco).

4 Three Kinds of Scope Objections

As reviewed above, Mitchell and Clement (1999) object to Barsalou's analysis of TRUTH, arguing that a "match" between the simulation of a visual event and a corresponding visual experience could represent not only *truth* but also *similar*, *comparable* and *looks like*. Adams & Campbell (1999) argue that MATCHING is neither necessary nor sufficient for TRUTH. Ohlsson (1999) objects that we would not be able to detect falsity from a mismatch. Siebel (1999) reminds us that there are situations in which a belief matches our perception of an event even though the belief may be false. Dove (2009) argues that experiencing voting cannot sufficiently track genuine instances of democracies.

On the surface, these objections all seem to raise similar worries, namely that modal symbols are insufficient to account for abstract concepts. However, once we distinguish between foundational semantic issues and issues concerning everyday categorization (between what determines content and what enables us to make everyday categorizations), we can see that they actually raise fundamentally different kinds of objections. Mitchell & Clement and Adam & Campbell essentially raise the semantic question of whether *truth* can be sufficiently represented in terms of a modal symbol. Dove raises the question of whether modal symbols can explain how we establish a referential connection to abstract properties (i.e., how we acquire abstract concepts), while Siebel and Ohlsson are mostly concerned with the empirical question of whether modal symbols can suffice to explain how we apply the concept of truth, i.e., whether they constitute our categorization devices.

I first discuss the semantic objections, arguing that they are shaped by semantic internalist intuitions that can easily be resisted. I then go on to discuss the other two objections pertaining to the questions of concept acquisition and categorization. I argue that for both, we can find relatively simple solutions that allow us to view even simplistic embodied and empiricist theories of abstract concepts (like the theory that TRUTH can be explained in terms of MATCHING) in a more positive light.

4.1 Semantic Objections

The reason Mitchell and Clement (1999) claim that MATCHING is not sufficient for a semantic account of TRUTH is that we can easily imagine cases in which two things match without them being true or false. Similarly, Adams & Campbell (1999) argue that we can easily imagine cases in which we would apply the concept of truth but not the concept of matching. In other words, representing the mental operation of matching is neither necessary nor sufficient to determine

the content of our ordinary concept of truth. Many involved in this debate have taken this argument to be a refutation of the proposal that MATCHING can explain TRUTH (see especially the commentary to Barsalou, 1999).

However, the issue is much more complicated. First, these arguments fail to distinguish between descriptive and foundational semantic issues of content. The objection can be interpreted as either arguing that the epistemic content of TRUTH is not covered by MATCHING alone, or that MATCHING cannot establish a sufficient referential connection to the property of being true. The former pertains to the descriptive problem of meaning, i.e., the question of what the meaning of a concept is, while the latter pertains to the foundational issue concerning content, i.e., the question of which facts determine that a concept has the content it has.

As a semantic objection (both descriptive and foundational), the above argument fails to be decisive even for such simplistic cases as explaining TRUTH in terms of MATCHING. The conclusion that MATCHING, for instance, does not suffice as the epistemic content of TRUTH or that MATCHING cannot determine the content of TRUTH presupposes a certain theory of content that is highly controversial. It seems that by arguing that MATCHING is neither necessary nor sufficient for TRUTH, Adams and Campbell presuppose that a successful account of content ought to specify necessary and sufficient conditions that can play the role of the epistemic content of a symbol and determine its semantic content (its reference). However, semantic accounts of concepts based on necessary and sufficient conditions (or *definitions*) have been called into question for very good reasons and need not be accepted by concept empiricists.[25] Moreover, prominent concept empiricists like Prinz (2002) or Barsalou (1999) explicitly reject this view both for an account of epistemic content and for the question of what determines reference.

To make even current concept empiricist proposals of abstract concepts appear more plausible, concept empiricists could embrace an alternative internalist view of content that is not based on necessary and jointly sufficient conditions. For example, empiricists could posit a number of sufficient conditions that determine the respective concept (some kind of *bundle theory* as proposed by Searle, 1958). Since MATCHING will probably be among these conditions, concept empiricists may not have given a complete semantic account of TRUTH but one that

[25] *Semantic definitionism* has become unpopular mostly due to Quine's (1957) arguments against the analytic-synthetic distinction and the lack of examples of successful definitions (Fodor, 1998; Prinz, 2002). But see Peacocke (1992) for such an account.

could at least be on the right track and that only needs to be completed once philosophers have agreed on an adequate account of the content of TRUTH.

Luckily, we do not have to wait for more elaborate theories of abstract concepts from concept empiricists. Barsalou for example has provided much more sophisticated explanations especially of TRUTH than is often acknowledged. His account goes far beyond the idea that TRUTH is reducible to MATCHING and relies instead on complex temporally extended simulations of external and internal events (Barsalou, 1999, p. 603). Similarly, according to Barsalou and Wiemer-Hastings (2005, pp. 136-137), abstract concepts involve representations of information extracted from events, the speaker and the listener. These additional conditions of TRUTH may not suffice as necessary and sufficient for truth but may still give a reasonably accurate description of the epistemic content of TRUTH.

Finally, concept empiricists could defend their view by rejecting semantic internalism, which has serious independent problems (e.g., classic Kripkean, 1972 or Fodorian, 1998 objections). To avoid these objections and also the above semantic scope objection, concept empiricists could turn to causal-historical foundational theories of meaning and argue that representing accurate descriptions, especially necessary and sufficient ones, are not necessary to establish reference. For example, concept empiricists could (and usually do, see Barsalou, 1999, Prinz, 2002) argue that the contents of TRUTH, ARGUMENT or DEMOCRACY may not be determined by their relation to other concepts but by a causal-historical connection to an actual property in the world, say a natural or social kind (Kripke, 1972; Putnam, 1975; Mallon, 2016; Khalidi, 2016).

Causal-historical relations could be mediated via modal symbols as proposed by Barsalou (1999) and to some degree Margolis (1998). Since the same set of concepts can establish a relation to several different concepts, VOTING may not only establish and sustain a causal relation to *voting*, but also to *democracy* and *freedom* (depending on the context). Similarly, contrary to Mitchell & Clement, it is not necessary that MATCHING only represents *truth*. Depending on the linguistic community, MATCHING can establish and sustain a referential relation to *matching*, *truth* and *similar*, just as we can imagine that the concept of liquid fluid can establish a relation to both water and twater (Kripke, 1972; Putnam, 1975).[26] None of this

[26] Causal-historical accounts of concepts have been applied to many concepts that are abstract, including natural kind terms like ‘atom’ (Putnam, 1975; Millikan, 1998) but also social kind terms like 'race' or 'sex' (Hacking,

suffices for a full explanation of how abstract concepts could be embodied. However, it already shows why the semantic scope objection is much more complex and much less convincing than admitted by some opponents of concept embodied cognition.

4.2 Objections Pertaining to Concept Acquisition

By embracing causal-historical semantics, concept empiricists' primary goal will be to show how we can perceptually establish and sustain a relation to kinds in the world that we cannot touch or see. According to Dove (2009; 2016), reaching this goal is unlikely. With respect to TRUTH, he doubts that a mental operation of matching can establish a connection to truth because MATCHING could also be triggered by instances of *matching*, *similar*, *identical*, *fit* and so forth. With respect to DEMOCRACY, he doubts that seeing people vote can establish a connection to democracy if nothing in our perception could tell us whether we are dealing with a genuine election.

This objection fails for the reason that causal-historical accounts of concept acquisition usually acknowledge that the same perceptual stimulus can in principle establish a causal-historical relation to several different properties or relations. For example, taking the classic example by Putnam (1975), the same perceptual features of water can establish a referential relation to both H2O and XYZ. Similarly, imagine that psychologists are right and we typically acquire and represent concepts by means of representing other concepts (sometimes called “features”) that are typical for the respective category (Rosch, 1975). Applied to abstract concepts like DEMOCRACY and TRUTH, concept empiricists could argue that the same experience can, depending on the context, establish a relation to truth or matching.

All it takes to be a concept empiricist with respect to concept acquisition is to argue that we always, typically, or sometimes acquire abstract concepts by means of modal symbols. This view is at least not obviously mistaken. We do usually acquire concepts (e.g., establish a referential relation to a property or relation) by having certain relevant experiences. In the case of TRUTH, suppose that a young child, by default, accepts every statement from their caregiver as true. This does not necessarily entail that they have a concept of truth, e.g., if the child does not yet conceptualize that statements can be false. At a certain point, the child however may

1999; Haslanger, 2005; Mallon, 2015). Such accounts are yet to be fully developed, but they are a highly promising alternative to internalist semantics and there is so far no reason to rule them out.

have acquired enough other concepts to notice a mismatch between what is being said and what is being perceived. One could say that these are the first steps towards acquiring "a feeling" for what it means for a proposition to be true even if this feeling may not yet be sufficient to constitute or determine the meaning of TRUTH.

However, once this initial step is taken, the child can collect further sensorimotor, introspective and emotional experiences about when and how to apply the concept of truth. In other words, they can now either establish a more reliable *sustaining mechanism* to TRUTH (see Fodor, 1998 or Margolis, 1998 for this terminology), if causal-historical accounts are true, or be able to explicitly or implicitly learn the definition of truth, if descriptivism is true. What these additional concepts are that allow the child to acquire the respective descriptions or sustaining mechanisms and whether they can be inferred from experience and represented in terms of modal symbols is ultimately an empirical question, but I do not see any a priori reason why this possibility should be excluded.

Similarly, imagine a child on election day. The child might hear many stories about the current government and the new candidates that they do not yet properly understand. Later they might witness their parents placing a piece of paper in a box. They may also experience situations in which they are able to speak their mind freely and instances in which this is not the case. All these experiences could be explained in terms of modal symbols that are combined in ways that eventually establish a reliable connection to a social kind of democracy (if causal-historical accounts are true) or that eventually allow the child to understand what essentially, or in amore bundle or family resemblance manner, constitutes a democracy (if descriptivism is true).

However, Dove might not be convinced. He might argue that the issue is not empirical at all. Whether perceptual symbols are reliable enough to establish a relation to *truth* or *democracy* is ultimately a conceptual question. Again, the main reason he questions the reliability of perceptual symbols in the case of TRUTH is that MATCHING could also be triggered by other properties like *similarity* or *identity*. With respect to democracy, he argues that seeing someone vote could not suffice to distinguish genuine from non-genuine instances of voting. We could witness false instances of voting and establish a connection not to democracy but to other concepts.

Both objections require a much more thorough discussion than provided by Dove (2009; 2011; 2016). So far, however, there are some initial reasons to resist his conclusion. First, none of his objections take the important roles of context and linguistic community into account, both of which have been crucial for causal-historical views of semantic content (e.g., Burge, 1979). As for the first objection, experiencing the mental operation of matching when playing Tetris will not track *truth*, just as seeing four furry legs and hearing someone bark in the wrong context does not tell us anything about dogs. What could explain how we can get from an experience of matching to *truth* is that we are often in the right contexts, e.g., when hearing someone making a statement.

Similarly, experiencing people vote in the wrong context, say if the voting process is merely a performance, might not establish a connection to something like a democracy. But why suppose that we are always or typically in the wrong context? I suggest that even children who do not live in a democracy will often be in the right contexts when experiencing instances of voting. These instances must not always be very sophisticated. Simply experiencing one's own family vote when choosing a movie, for example, may suffice. Finally, why should it matter whether the voting actually constitutes a democratic election? Just as acquiring the concept of a crocodile does not require us to actually see a real crocodile, the mere appearance of a democratic election or the description of one in a book may suffice to acquire this concept. As long as we can rely on our linguistic community, fake elections may just as well establish a relation to *democracy* as actual democratic elections. Perhaps in some cases this connection might go wrong and the citizens of a dictatorship, when using the word 'democracy', do mean something that we would not call a democracy. However, Dove has not shown that this should always or even typically be the case.

Moreover, the same concerns could be raised against the perceptual acquisition of concrete concepts. Consider, for instance, a case in which a child acquires the concept of dog merely by means of seeing illustrations of dogs. According to causal-historical views, this is possible because the pictures refer to a genuine kind in nature and because there is a causal-historical link between dogs, the illustrator and the society the illustrator and the child are part of. According to Dove, this works for DOG because dogs can more reliably be tracked by their appearance. This, he argues, is more difficult for abstract concepts like DEMOCRACY where the connection is “loose at best” (p. 419) and because “little direct connection exists between these perceptual features and what makes a government a democracy” (ibid.).

First, the problematic metaphysical picture drawn by this argument is not acknowledged by Dove. The idea seems to be that having a certain shape and being able to bark is connected more closely to whatever constitutes being a dog, e.g., a certain history or homeostasis producing mechanism (e.g., Boyd, 1999), than the perceptual characteristics of a democracy. Even if this were the case (Dove does not properly argue for this assumption), why conclude that connection is therefore *too* loose to track democracies? Whatever it is that makes it the case that a country is a democracy, perhaps certain attitudes, duties and rights (Searle, 2010; Mallon, 2016), why suppose that it cannot reliably be tracked by certain superficial characteristics, such as seeing people vote or speak freely?

One could go even further and argue that the connection is stronger at least in some cases. Having a certain shape and being able to bark does not constitute being a dog – it is only evidence for there being a dog. Being able to vote and speak freely may arguably constitute living in a democracy if this is literally what we mean by living in a democracy (again this is a difficult ontological question not addressed by Dove). Similarly, experiencing events of voting and its immediate effects, for instance, when choosing a film in one's preferred social group may be exactly what it means to be in a democracy (and may not just be evidence for it).

Another way in which concept empiricists could respond to Dove is by arguing that fake instances of voting only look like true instances of voting because in the majority of cases voting does track democracies. This case is similar to more familiar cases, e.g., in which we mistake a crumpled bag for a dog (Fodor, 1998). The question that is raised in these instances is: what makes it the case that our concept of dog refers to *dog* and not to *dog and crumpled bag in the dark* if they all cause DOG?

Fodor's (1998) answer to this problem is that DOG refers to dogs and not dogs and crumpled bags because the causal connection in this case is asymmetrical. This means that crumpled bags only trigger DOG because *dogs* usually trigger DOG. If dogs did not usually trigger DOG, crumpled bags that look like dogs would not trigger DOG either. Moreover, DOG is triggered in many circumstances in which CRUMPLED BAG is not triggered and CRUMPLED BAG is not triggered in most situations in which we see dogs. Applied to the concept of DEMOCRACY we can say that VOTING might be triggered by undemocratic elections only because VOTING is usually triggered by democratic ones. If all elections were undemocratic they would be of no

use to dictators who want to legitimize their government. Moreover, not all instances of voting trigger the concept of democracy. Again, the context is crucial. Seeing a performance of an election might not trigger DEMOCRACY simply because the context is not right, which does not mean that voting in general cannot track democracies.

Finally, some philosophers of language have argued that merely knowing a word suffices for concept acquisition (see e.g., Millikan, 2017 or Burge, 1979). Note that this is not just an outlandish view in philosophy, but instead, even though controversial, based on a theory of content that is extremely popular (see also, Millikan, 1998 and Fodor, 1998). Since words, it has been argued, could be represented by means of sounds and visually represented letters, hence in terms of modal symbols, it is at least conceptually possible that we acquire abstract concepts by means of acquiring the right linguistic modal symbols, assuming again we are in the right context and our signs are embedded in the right linguistic community.

Again, none of this suffices as a fully worked out solution to the problems raised by Dove. However, it does suggest many promising options that are available to concept empiricists and proponents of the embodied cognition hypothesis. Importantly, it shows that it is simply not obvious that even the most abstract concepts cannot be acquired by means of modal symbols. For instance, even concepts like LIVING THING or CHILIAGON (a polygon with 1000 sides) could, thus, be acquired simply by learning the word ‘chiliagon’. Further work needs to be done to establish a better picture of the nature of abstract concepts and the relation between different accounts of concepts before we can draw Dove's very strong pluralist conclusions.

4.3 Objections Pertaining to Concept Application

Finally, a third issue that concept empiricists have been criticized for is their account of concept application. Siebel (1999), for example, argues that if Barsalou were right, we would always apply the concept of TRUTH whenever two things match. Similarly, Ohlsson (1999) worries that we would not be able to detect falsity from a mismatch. These objections seem to tackle neither the semantic problem of how to determine the content of a symbol nor the problem of how we acquire the respective concept. Instead they raise the question of whether concept empiricists can explain how we *actually* apply an abstract concept like TRUTH to the world, which presupposes that we already possess it.

Once the distinction between semantic accounts of concepts and accounts of that which enable us to apply a concept (what I called our *categorization device*) is in place, we can find relatively simple solutions to the problems raised by Siebel and Ohlsson. First, there is no reason why the same set of beliefs (or categorization device) could not be employed for several distinct categorizations even pertaining to the same object. For example, the mental process of matching may enable us to assign not only the concept of truth but also the concepts of identity and similarity to the same kind of thing (e.g., a statement). This does not speak against concept empiricism if this is in fact how we use the concept of matching or how our respective categorization devices are constituted.

Secondly, the same concept can be used to subsume different objects under different concepts. Just as the concept of furry can be used to identify not only fur but also cats and dogs, the concept of matching could be used to identify true statements or the missing piece of a jigsaw puzzle. Context may suffice to allow us to make the classifications that are adequate in the respective situation. Siebel, however, seems to deny that this is plausible and claims that, according to concept empiricists, we would always apply the concept of truth whenever we apply the concept of matching. However, I cannot see any reason why concept empiricists should be committed to this claim. Again, concept empiricism could only be ruled out if it produced absurd predictions, but this is, at least with respect to the application of the concept of TRUTH by means of MATCHING, not the case.

Contrary to Siebel and Ohlsson, I suggest that one of concept empiricists' major strengths is their account of concept application. It is at least at first glance in accordance with common sense that we tend to subsume objects under concepts by means of information that is perceptually available. This is the case for concrete concepts like WATER, just as it is the case for abstract concepts like DEMOCRACY. We do not normally identify water as water by means of its molecular structure. Instead we classify the liquid in front of us as water based on its appearance. Similarly, it is not unreasonable to assume that we apply concepts like TRUTH or DEMOCRACY based on what we can see or feel. For example, in order to categorize a country as a democracy we want to see whether it enables citizens to vote and speak freely. This does not mean that we believe that merely seeing people put paper in a box is enough to proof that a country is a democracy. If we have doubts, we need to gather more information on the voting process. None of the commentators have been able to make a convincing argument for the idea that this could not also be accomplished by means of modal symbols.

In the case of TRUTH, we can hypothesize that the feeling of a match of proposition and perception is at least among the typical criteria we rely on to categorize a statement as true. Think of cases in which we classify a statement as false simply because it doesn't "ring true" or "feel right", even if we cannot exactly say why. In other instances, we may go through intense investigations by gathering evidence. This is, for example, the case when seeing an illusion. Recall that Siebel (1999) doubts that we would be able to distinguish illusions or hallucinations from real instances of truth if TRUTH were reduced to MATCHING. Again, nothing commits concept empiricists to such a one criterion view as long as the other concepts used to identify illusions can be explained by means of modal symbols. Again, so far, no arguments have been provided that show that they could not.

5 Conclusion

I showed that many of the objections to concept empiricism and embodied cognition addressing the scope problem conflate issues pertaining to content with issues pertaining to concept application and concept acquisition. I argued that contemporary concept empiricism (unlike traditional or logical empiricism) and especially the embodied cognition hypothesis are theories primarily of concept application and that especially the objections to empiricist accounts pertaining to this issue are extremely weak. It is plausible to limit concept empiricism to concept application, but even if applied to the other two issues (content and concept acquisition), concept empiricists' accounts of abstract concepts are much stronger than is often acknowledged. In particular, I showed that some of the semantic arguments raised against concept empiricism presuppose problematic descriptivist and definitionist accounts of content.

None of this shows that concept empiricism or the embodied cognition hypothesis are correct. However, I was able to show that a wide-scope concept empiricism is at least empirically possible. In other words, I take it that at least in principle nothing speaks against the hypothesis that even highly abstract concepts can be acquired and applied by means of modal symbols. Since there is in fact some evidence for abstract embodied cognition I take it that this is not only a theoretical possibility. However, since there is also much evidence for modal symbols (Machery, 2016), I predict that a pluralist account might still be successful. Importantly, such a pluralist account may not cut across types of concepts (abstract or concrete), but instead apply to all concepts. I thus speculate that both abstract and concrete concepts may be represented by both modal and amodal symbols.

Chapter 4: Social Constructionism, Concept Acquisition and the Mismatch Problem

Abstract: An explanation of how we acquire concepts of kinds if they are socially constructed (e.g., *man* or *bachelor*) is a desideratum both for a successful account of concept acquisition and a successful account of social constructionism. Both face the so-called "mismatch problem" that is based on the observation that that there is often a mismatch between the descriptions proficient speakers associate with a word and the properties that its referents have in common. I argue that externalist theories of reference provide a plausible and attractive account of concept acquisition, including the acquisition of concepts of social constructs, that avoids the mismatch problem. However, externalist theories are ontologically and psychologically highly demanding, which places strong constraints on accounts of the metaphysics of socially constructed kinds. In particular, they require a rather strong form of realism that is incompatible with some but not all theories of social constructionism. Finally, I show that these demands can be met by means of adopting a homeostatic property cluster view of natural kinds.

1 Introduction

Many of the categories we care most deeply about are socially constructed, i.e., they do not capture kinds that could be discovered and studied by physics, chemistry or biology, but kinds that constitutively rely on contingent facts about our social relations (Diaz-León, 2015, Mallon, 2014). Kinds that are clearly socially constructed include *president of the United States of America*, *bachelor, money* and *tax payer*. More controversial examples include *race*, *gender, age* and *sex*.[27]

[27] I use capital letters to denote concepts, single quotation marks to denote lexical expressions, italics to denote properties and double quotation marks to denote sentences and technical terms.

To be able to think that the current president of the United States is lying or worry that there will be another recession is to take an attitude towards a proposition (e.g., the proposition that the current president of the United States is lying). The constituents of these propositional attitudes are called "concepts" (Rey, 1983; Peacocke, 1992; Fodor, 1998; Margolis, 1998; Löhr, 2018). Concepts have contents, i.e., they are about objects, properties or relations in the world. Mental content (at least partly construed as reference), i.e., what a certain concept is about, is determined by means of either internally represented inferential relations or descriptions and/or external causal-historical relations between the concept and a property or an extension. According to "semantic internalism", mental content is exclusively determined by internal mental representations. According to "semantic externalism" mental content is determined, at least partly, and at least in some cases, externally.

So-called "foundational theories of content" (Lewis, 1970), i.e., different versions of semantic internalism and semantic externalism, are fundamental for any theory of concept acquisition because they determine (at least in part) a concept's possession conditions. A person has acquired a certain concept if and only if she meets its respective possession conditions. So, if the referent of a concept is determined by an internally represented description, then, to be able to refer to this referent and to possess this concept, one must represent this description. If reference is determined by, e.g., a causal-historical relation to an external entity, then to possess the respective concept, one needs to be part of the right causal chain. As a consequence, according to semantic externalists, one need not necessarily know the description or inferential relations that unambiguously pick out the relevant reference (e.g., one does not need to hold many true beliefs about the respective kind) to possess the respective concept.

The vast majority of research on concepts and their contents, both in philosophy and psychology, has focused on chemical kinds like *water* or *gold*, biological kinds like *bird* or *dog* and functional or artificial kinds like *table* or *chair*. Seldom have these theories been applied to more abstract kinds, especially social kinds like *gender* or *art*. The acquisition of concepts of these latter kinds is difficult to explain by both internalist and externalist views of concept possession.

At first glance, internalist approaches to mental content and concept acquisition appear to fare better with respect to socially constructed kinds than externalist alternatives. While semantic externalism appears to require a real kind that the subject can be causally related to, internally

represented inferential relations or descriptions can pick out a property even if it is not instantiated, is unstable, mind-dependent, or not directly accessible to perception.

However, semantic internalists have difficulties explaining how we could be systematically or collectively wrong about the properties that define the reference of our ordinary concepts considering that they argue that knowing these properties is required to refer to the respective property. For instance, categories like 'race' and 'gender' have often been assumed to capture biological kinds, when, in fact, at least the majority of today's biologists and social constructionists agree that the properties we associate with both terms cannot be fully explained by biological properties. If internalism were right, this would mean that our concepts pertaining to race and gender do not refer. At least according to some social ontologists, this is counter-intuitive. For example, we generally do not have the impression that sociologists who study gender and race are talking about something that does not at all exist, even though some philosophers have made such an argument.[28]

This problem has been called the "mismatch problem" (Glasgow, 2009; Mallon, 2017). It refers to the mismatch between the properties that at least lay people take to be essential to a certain kind and the actual common properties of the kind. Solving this problem is a crucial desideratum not only of a successful theory of concept acquisition, but also of a successful defense of social constructionism. In other words, it is difficult to see how theories of concept acquisition and social constructionism that cannot explain how we can think about socially constructed kinds could be successful.

Semantic externalism seems to fare much better when it comes to avoiding the mismatch problem. According to externalists, one can refer and think about an entity as having a certain property even if one lacks the descriptions that unambiguously identify the respective property. This would mean that one can, for instance, think and make predictions about white men even if one were to believe, according to social constructionists falsely, that 'white man' refers to a biological kind.

However, semantic externalism comes at a price that especially social constructionists may find difficult to pay. The first problem is that semantic externalism requires a rather strong form of metaphysical realism. It not only requires that the respective kind to which one is causally

[28] It might of course be that even if our lay concepts of social kinds did not refer, the concepts of experts could refer, nonetheless. However, it is not clear which these referring expert concepts are considering the immense disagreement in the respective literatures in the social sciences.

related actually exists (which at least many social constructionists accept), but that it exists independently of the minds that represent this property. Furthermore, it requires that the kind must be stable in a way that allows it to be tracked by our minds. Both conditions seem to be denied by many social constructionists for social constructs. According to them, socially constructed kinds are essentially mind-dependent and highly unstable (e.g., Searle, 1995; Hacking, 1999). [29]

The second problem for externalism of concept acquisition is that it requires some psychological commitments in order to explain how individuals can establish a relation to an external kind. An answer to this problem should address the so-called "qua problem" (Devitt, 1981), i.e., the problem of discovering the grounds that fix the reference of an expression or concept. A version of this problem in connection with the mismatch problem has recently been addressed by Mallon (2017) and will be discussed below in section 4.

In order to approach the qua problem it is often assumed that we have to posit a number of innate biases, such as a whole-object bias or a basic-level bias, that explain how the child goes from the superficial regularities it encounters to track the relevant kinds, while ignoring the irrelevant ones. For example, it is generally assumed that we need some innate biases to explain how we track *water* and not *water and its surroundings* or how we acquire the concept of ball as opposed to merely the concept of football. A similar problem occurs for social kind concepts like the concept of mother or the concept of naughty that the child seems to acquire very early in its development (Tardif et al., 2008). It is not clear, however, whether the constraints that could potentially solve the qua problem for physical kinds can also explain how we acquire concepts of more abstract social kinds.

So, the main approaches to concept possession (semantic internalism and semantic externalism) are difficult to apply to socially constructed kinds. This puts severe pressure both on a theory of concept acquisition and a theory of social constructionism. On the one hand, if our theories of concept acquisition cannot explain how we can think about social constructs, these theories must be false (under the assumption that we do, in fact, think about social constructs). On the other hand, if we cannot acquire concepts of social constructions, and consequently cannot think about them, social constructionism must be false (since any plausible theory of social constructionism is committed to the assumption that we do, in fact, think about social constructs).

[29] I introduce different notions of mind-dependence below.

In section 2, I introduce Margolis's (1998) externalist account of concept acquisition for biological kinds. In section 3, I explain why such a view is both ontologically and psychologically very demanding. In section 4, I discuss in more detail in which way exactly an externalist view of concept acquisition demands that our referents need to be real and mind-independent in a very strong sense. In section 5, I show that social constructionists can meet these demands by endorsing a homeostatic property cluster view of socially constructed kinds. In section 6, I review some evidence for innate social biases that may be employed to tackle the qua problem for social kind concepts.

2 How to acquire a biological kind concept

One of the main challenges for any successful theory of concept acquisition is to show how we can get from the superficial properties of a kind that we can directly perceive to the kind itself (e.g., how we get from *transparent liquid* to *H2O*). Internalists about concept acquisition argue that this is only possible if we, at least implicitly, represent sufficiently many true descriptions or inferential relations that unambiguously pick out the concept's referent (Searle, 1958; Peacocke, 1992). Consequently, according to internalists, concept acquisition can be extremely effortful. To be able to refer to (and, consequently, think about) *water*, for instance, we have to engage in much scientific research. In other words, internalists tend to argue that before we knew that water is essentially H2O we were not able to think about water, but only about drinkable tasteless transparent liquids.

Externalists about concept acquisition (Fodor, 1998; Margolis, 1998) argue that we can get more directly from superficial properties to the respective kind, as long as both are reliably correlated and this correlation can be explained by an underlying common essence or mechanism. For instance, they argue that we can think about *water* if our concept of water is reliably triggered or caused by a set of superficial properties like *transparency, drinkability* and *liquidity* whose co-occurrence can be explained by a common mechanism or essence, e.g., its molecular structure. This inductive behavior of tracking a kind by means of its superficial properties is what Margolis (1998) refers to when he argues that we can acquire concepts by means of detecting their referents via their symptoms.[30]

[30] Since we have to internally represent what these symptoms are one might argue that Margolis defends a hybrid view of mental content (as, proposed by Evans, 1973 or Devitt, 1981). However, since these representations are not descriptive, i.e., they do not determine the reference of the respective concept (what Recanati, 2012, calls a "non-descriptive mode of presentation"), in my view, a necessary condition for hybridity is not met.

In addition to acquiring a concept by means of its referent's perceptual features or symptoms, Margolis argues that especially experts can, and usually do, rely on more theoretical beliefs to track kinds. To use Margolis' example, some experts have enough theoretical knowledge and the necessary tools to detect the presence of a proton in order to establish a referential relation to protons even if they cannot directly see them (see also Kripke, 1972 and Putnam, 1975). Less informed people can rely on the knowledge of these experts in order to refer to external objects that they do not have direct perceptual access to. For example, externalists argue that one can acquire the concept of arthritis even if one has mistaken beliefs about what 'arthritis' refers to as long as one's use of the word is embedded in an expert community that uses it to refer to an inflammation of joints (Burge, 1979; Millikan, 2017).

It is usually assumed that the referential tracking of kinds by means of their superficial properties works especially well for chemical natural kinds like *gold* or *water* (Kripke, 1972; Putnam, 1975), which are usually thought of as mind-independent and stable, i.e., what Hacking (1999) calls "indifferent kinds" (e.g., water retains its molecular structure independent of our mental states). Mind-independence is necessary for an externalist account of concept acquisition because we can only establish an external causal relation mentally to a kind if it does not need our minds for its existence (see next section). Stability (Rey, 1983; Löhr, 2018) is necessary because it is not clear how semantic externalism could be true if our referents and especially the relation between them and their co-occurring properties would constantly change. For example, if the essence or nature of water as well as its superficial properties would constantly change in unpredictable ways, it is not clear how we could ever establish a reliable external relation to this kind.

In addition to the world offering a stable and mind-independent source of contents, Margolis (1998) argues that externalist accounts of concept acquisition are also psychologically demanding. First, we need to explain why we tend to go from superficial features to a common underlying property in the first place (i.e., why we do not simply talk about superficial regularities). To explain this, Margolis cites psychological studies suggesting an innate bias to assume an underlying hidden essence as the mechanism that produces regularities in the environment, which is called "psychological essentialism" (Medin & Ortony 1989; Machery, 2014). For example, we usually detect a raccoon by its appearance and typical behavior, but this does not mean that we think that raccoons are defined by these properties (Keil, 1989).

Secondly, Margolis (1998) argues that we need to assume further fundamental cognitive abilities in order to explain the kinds of categorizations human beings find most natural. Why, for instance, do children infer correctly that an ostensive definition of *ball* refers to the ball alone and not the ball and its surroundings or only a part of the ball? Following research by Soja, Carey, & Spelke (1991) and Carey (2009), Margolis (1998) argues that we have several cognitive biases, e.g., towards whole objects, that explain why children usually find whole objects more salient than objects and their proximate surroundings.

By combining semantic externalism with internal psychological biases, externalist views of semantic content and concept acquisition, like the one proposed by Margolis, are especially well-equipped to avoid both the mismatch problem as well as the qua problem. They can avoid the mismatch problem because they do not require the possession of many (or any) true beliefs about the properties of the reference of our concepts. They potentially avoid the qua problem because they allow for at least some internal representations that promise access to a correlated single real kind, without determining what this referent is. For example, the reason that the child, when interacting with dogs, acquires the basic-level concept of dog before it acquires the superordinate concept of animal (Lakoff, 1987) and not the concept DOG AT TIME 1 AND CAT AT TIME 2 is that it has a disposition to find whole objects salient (due to an innate whole object bias) and because the superficial properties of dogs it encounters are strongly correlated with the real kind *dog*, but not the kind *dog at time 1 and cat at time 2*.

3 Constraints on a theory of concept acquisition

Externalist theories of concept acquisition, such as the one proposed by Margolis (1998) offer a plausible and attractive theory of how we acquire concepts of chemical and biological kinds. They are plausible because they require neither too much nor too little knowledge for concept possession and the knowledge they do require (for contingent reasons) is psychologically plausible (there is empirical evidence that children possess it). They are attractive because they avoid the mismatch problem while also, at least potentially, avoiding the qua problem. In other words, they assume enough psychological capacities to avoid the qua problem, but deny that these representations are necessary and sufficient to determine the reference of the concept, thereby avoiding the mismatch problem.

However, the same features of Margolis' account of concept acquisition that make it resilient against the mismatch problem and the qua problem also make the view ontologically and psychologically highly demanding. First, the account requires that the world is already divided

into real "inductive- or scientific kinds" (Boyd, 1999) before we can think about them (which is a requirement for solving both the mismatch problem and the qua problem). Secondly, it requires that we have some means of detecting these kinds. This means that their properties must be, at least to some degree, perceptually accessible (which is a requirement for an empirically plausible explanation of how we actually acquire the respective concept). Thirdly, it requires that the individual comes equipped with several cognitive biases and inductive capabilities in order to explain, for instance, why we tend to think about real underlying kinds and not mere superficial regularities (which is important to explain how we actually acquire concepts and to avoid the qua problem).

These requirements put strong constraints on any theory of the metaphysics of socially constructed kinds. If social constructionists wish to apply externalist theories of reference to avoid the mismatch problem they need to accept that social constructs must be real kinds that are mind-independent, stable and perceptually accessible. Furthermore, to avoid the qua problem, they must accept some internal representations that constrain the range of possible referents. However, if externalists of concept acquisition aim to explain how we can acquire not just chemical and biological kind concepts but a wide range of concepts, including concepts of socially constructed kinds (what Prinz, 2002 and Löhr, 2019 call "the scope requirement"), they too should hope that constructed kinds meet these requirements.

At least the first requirement of mind-independence and stability, however, seems incompatible with most current social constructionist theories, according to which many social kinds constitutively dependent on mental states (e.g., Searle, 1995; Hacking, 1999, Mallon, 2004). Thus, in the next section I take a closer look at what exactly it means for a social construct to be real, mind-independent and stable in the sense relevant for the applicability of externalist theories of concept acquisition. I argue that the mind-independence that is required is highly demanding. In section 5, I argue that this demanding mind-independence, as well as the demands of stability and perceptual accessibility, can be met by endorsing a homeostatic property cluster view of socially constructed kinds. In section 6, I review some evidence for biases relevant for tracking social kinds.

4 What kind of realism do we need?

To give a more detailed account of the kind of realism that is relevant in the present case (and to show how strong this requirement actually is), I first introduce a distinction between social constructs that already exist and that can already be detected and social constructs that are yet

to come into existence. Secondly, in the set of already existing social constructs, I would like to distinguish the constructs for which experts already have sufficiently many true descriptions from the constructs for which we currently lack expertise. For instance, *bachelor* is a social construct that already exists and whose defining features are, at least arguably, relatively uncontroversial (a bachelor is an unmarried adult male). *Man*, on the other hand, is an already existing kind whose defining features are currently debated.

For externalist accounts of reference to be applicable to socially constructed kinds, it is sufficient that these kinds already exist and that experts agree on a description that reliably tracks the respective reference. Even though non-experts may lack this description, they can refer to the respective kind by deferring to these experts (Burge, 1979). A deeper problem surfaces in cases where there is no expert community available (contrary to the case of *arthritis*) and in which no stipulation is sufficient (contrary to, arguably, the case of *parent* or *bachelor*, see Haslanger, 2005). This is the case especially for so-called "covert social constructs" (Mallon, 2017), e.g., those pertaining to race, sex or gender, which are exactly the kind of concepts to which the mismatch problem is especially applicable. For these kinds, there is still much debate on what the underlying causes of the correlated properties are, even among experts. I argue that these kinds have to be mind-independent and stable in a very strong sense. This means that, as in the case of *racism* and *recession*, the existence of these kinds must not depend on us thinking of instances as having a certain race or gender (see Khalidi, 2015 for a categorization of different kinds of social kinds). This does not mean that these kinds do not depend on any mental states at all (I agree that the social construct *white man* constitutively depends on our mental states). However, the mental states in question cannot be the ones that contain the respective social kind concept as a constituent, as this would be circular. In other words, I argue that it is conceptually not possible (it would be circular) to turn to semantic externalism (in order to avoid the mismatch problem and give a promising theory of concept acquisition) and also hold on to the claim that a particular social kind is constituted by the same mental state that has the respective property as its content.

The reason, in a nutshell, for this rather strong covert kind realism is that, by accepting that the relevant notion of concept is defined as that which constitutes our propositional attitudes (Fodor, 1998; Margolis, 1998), we cannot think, in principle, about something *as* something unless we already have the concept of it. We cannot think of someone as being heterosexual, for instance, unless we already have the concept of heterosexual. According to externalist accounts of conceptual content, to be able to think about something as having a certain property

means to be in the right causal relation to this property. In the case of *heterosexual,* this means that we only possess the corresponding concept if we are in a referential relation with the kind *heterosexual* (e.g., via its instantiations). This requires that this property must have existed before we had the thought about someone as being heterosexual (as we can only relate to something that exists).[31]

Compare this case with the social kind *money*. According to Searle (1995), we have to think about certain concrete entities (e.g., pieces of fabric) *as* money in order for these entities to have any monetary value, i.e., for money to exist. Only if enough people have certain beliefs about the value of bitcoins can a bitcoin have any value at all. So, the first person who has ever had a thought consisting of MONEY (referring to the kind that exists today) must have represented either the description that picks out *money* and only *money* or must have relied on *money* as an already established kind. In the latter case, *money* could thus not have been established in the way proposed for instance by Searle (1995), i.e., by means of mental states that constitute this concept. Again, the internalist picture works well for invented kinds like *bitcoin* (it makes sense to assume that somebody had the proper description of bitcoins that simultaneously brought them into existence), but not for covert social kinds about which we lack expertise and that were not really invented in the same sense as bitcoins were invented (*racism* or *white man* rather seemed to have developed from complex social relations).

Ron Mallon (2017) recently put forth a different solution to both the mismatch problem and the qua problem. He argues that we can rely on Evans' (1973) notion of "referent switching" in order to avoid both problems. For instance, he argues that in the case of 'white person', we could have first had a label that distinguished people purely based on their skin color (what he calls a "weak natural kind"). Later, the referent could have switched to the social kind that we today associate with the expression 'white person' (at least according to social constructionists). This solution is adequate for what we could call a "linguistic mismatch problem" (the problem that occurs if the descriptions we associate with a word do not match the actual properties of the kind), but not for the more basic and more problematic "conceptual mismatch problem". It is clear that we can change the meaning of our terms by attaching new concepts to the same terms, but it is not clear how we manage to acquire these new concepts in the first place. The deeper conceptual mismatch problem is the one that makes the mismatch problem so difficult.

[31] Simultaneous construction, too, is not compatible with semantic externalism. Imagine by simply labelling a group of people (e.g., "the leaders") we invent a social kind. In this case we do not have to do with a covert kind because the application conditions of this kind are known to us (because we invented the category).

Moreover, a solution to the conceptual mismatch problem can be viewed as a solution to the linguistic mismatch problem.

Mallon's reference switching account is thus a) too weak for the conceptual mismatch problem and b) not required. It is too weak because it does not explain how we have acquired the new concept that we "switched" to in the first place (again in the externalist framework that Mallon seems to endorse we cannot simply change our descriptions, but need to causally relate to another kind). It is not required because once we acquired a new concept – in this case the concept referring to the socially constructed kind, as opposed to the merely weak natural kind – all that is left to do is the philosophically rather trivial (and only psychologically demanding) task of associating this new concept with the already familiar linguistic label. Again, what is crucial is not this task of concept switching but that of acquiring a new concept of the relevant kind. This kind (for reasons essential to the semantic externalism of covert kinds) must have existed before we first used it in thought (for instance to associate a familiar lexical sign with it). Moreover, it must have existed in a stable manner so that we can detect it by means of its superficial or theoretically accessible deeper properties.[32]

If *woman* or *white* are covert kinds, as argued by many social constructionists, i.e., if the mismatch problem applies to these kinds in the strong way in which currently nobody knows what the right description of these terms is, it follows that *woman* and *white* must be real mind-independent kinds. They must be real kinds in the same strong sense in which *racism* and *recession* are real mind-independent kinds (few people would argue that racism only exists because we think it exists)[33]. Note, again, that we needed a causal historical account to explain the mismatch problem for exactly these cases (especially those pertaining to *gender* and *race*). Thus, it is at least a necessary condition for a non-descriptivist account of covert social constructs that these constructs must have existed *before* the first person was able to think about

[32] To be absolutely clear, none of this means that we cannot invent any social kinds (we can arguably simply invent social kinds like *president* or decide that by "parent" we only mean *primary caregiver*). However, these are the easy cases for which the mismatch problem does not arise in the same way in which it arises to genuine covert kinds like *white man* for example. Moreover, overt kinds are not the kind of social kinds that we need causal-historical accounts of reference for. The kinds that lead to the mismatch problem, i.e., the kinds that make social constructionism especially interesting, requires a strong kind of realism. In other words, if *woman* is the covert kind that social constructionists argue it is, then it cannot be a kind that was invented in the same way that we invented *blog* or that we decided what we mean by "parent". *Man, cis* or *heterosexual* must be more like *racism, mansplaining* or *recession*, i.e., real kinds that were discovered and then named, as opposed to invented by naming it.

[33] Racism and recessions exist because there are minds. They are thus mind-dependent in the weak sense that they exist because there are minds. However, we do not need to think about things as racist or as a recession in order for them being racist or a recession. In this stronger sense, both kinds are not mind-dependent.

them (i.e., form beliefs with concepts of these kinds). Reference-switching is at play in this process, but it depends on the much more fundamental problem of acquiring the right concepts that we can switch to.

5 How to Acquire a Concept of a Social Construct

I have argued that if we turn to semantic externalism for an account of concept acquisition that avoids both the conceptual as well as the linguistic mismatch problem we have to assume that covert social-, chemical- and biological kinds are similar at least in the way that they are all stable and mind-independent in the relevant sense. Moreover, we have to assume that there are some psychological mechanisms that allow us to detect these real social constructs in order to avoid the qua problem and to explain how we, in fact, establish a referential relation to socially constructed kinds. At first, these requirements appear incompatible with core claims of social constructionism. This is because social constructionism seems to be committed to the idea that many social kinds, especially those pertaining to sex, gender and race, are essentially mind-dependent and unstable (see again Searle, 1995; Hacking, 1999; Mallon, 2004). Moreover, since social constructs are relatively abstract, we might wonder how we can establish a direct relation to them and whether there could be psychological biases in place that could help us track only relevant kinds. In this section, I argue that there are good reasons that at least covert kinds like *woman* or *white* meet the conditions of stability, perceptual accessibility and mind-independence. In the final section, I review evidence for relevant social biases.

5.1 Essences

The main reason why it may be problematic to think of social constructs as real mind-independent kinds in the sense specified in the previous section is that, according to social constructionists, social kinds like *man* or *white* (unlike *water*) do not have essences (in the sense that water is essentially H2O). For example, one of the key assumptions that many social constructionists share is that the properties that are usually associated with, say, men, cannot be explained by a biological property that all men share (Y chromosome, for instance). Instead, they insist that these generalizations ought to be explained by social properties or mechanisms (Mallon, 2003; Diaz-Leon, 2015). Moreover, it is widely agreed upon that social kinds lack essences altogether and that social sciences, unlike, perhaps, physics or chemistry, are not in the business of discovering essences. If mind-independence depends on essences, social constructionism seems incompatible with semantic externalism (which requires this mind-

independence) and, thus, social constructionists cannot turn to externalism to solve the mismatch and the qua problems.

Fortunately, the view that natural kinds are individuated by means of essences is highly controversial and rejected by many contemporary philosophers of science. A more popular theory of natural kinds is Richard Boyd's (1999) influential homeostatic property-cluster view (HPC). According to HPC, a natural or scientific kind may be characterized as a cluster of more or less co-instantiated properties and some mechanism that explains this co-occurrence. Because this mechanism does not necessarily have to be based on a molecular structure, as in the case of water, but can be a social mechanism, such as common beliefs or institutional and historical contingencies, Boyd's account of natural kinds can be and has been applied to social kinds. Applied to gender, for example, Boyd (1999) and Mallon (2003) argue that a set of social practices and beliefs about women and men constitute the social roles of *woman* and *men.* These social roles can function as mechanisms that explain certain regularities of behavior in people who are labelled accordingly. For instance, it could explain why individual men often exhibit gendered behavior and traits.

Importantly, the HPC account allows that covert social kinds may not be dependent on the mental states that consist of concepts picking out this kind, but on different social practices, institutions and mental states that are realized by or based on mental states with other contents. This can explain how kinds like *woman* or *heterosexual* could be more like *racism* or *recession* in the sense that they could be discovered (rather than invented in the standard sense of 'invented'). They could be discovered even if they constitutively depend on our mental states, i.e., even if *women* or *man* would not exist without the existence of mental states. Kinds that can be discovered in this way are real and mind-independent enough to meet the first condition for semantic externalism specified above.

5.2 Stability

A second reason why we may not be able to apply externalist accounts of concept acquisition to social kinds is that their mind-dependence may render them too unstable to be referred to in thought. So, while it may be that our covert social kinds are real and mind-independent in the sense spelled out in the previous subsection, this mind-independence may not be stable enough to be applicable for causal theories of reference. However, relying again on the HPC account of scientific or inductive kinds, we have good reasons to assume that the mechanisms that can

explain the properties associated with many social kinds are stable enough to sustain a reliable external referential relation.

First, there is no reason to assume that social constructs are so unstable that we are not able to think about them at all even if the reference changes over time. A slowly changing kind may still remain stable long enough (usually even too long from the perspective of politically-engaged social constructionists) for us to track it.

Secondly, stability can be achieved by means of so-called "looping effects" i.e., the same phenomenon associated with social kinds that has led some theorists to posit a strict distinction between natural and social kinds (Hacking, 1999). Once a social kind is established and people are classified as such, these individuals may adjust their behavior and mental states accordingly, which further perpetuates and stabilizes the existence of this social kind (for example, a student labelled as unintelligent may lose their motivation to study). Moreover, it might be difficult for individuals to escape being constructed in a certain way if social constructionists are right and one's community has a constitutive impact on one's personality and behavior (Burr, 2015). Finally, psychological biases may also stabilize a social construct. For instance, if one is born into a social group with a bad reputation, it might be difficult to be accepted by members of another social group due to prejudices and other (often innate) in-outgroup biases (Mullen et al., 1992).

5.3 Perceptual access

So, there are several reasons to think that covert social kinds are stable and mind-independent enough to be compatible with externalist accounts of concept acquisition. However, a third reason why we may not apply externalist accounts of concept acquisition to social constructs is that it seems that we have no perceptual access to social kinds because they are too abstract. This not only threatens the idea that externalist accounts of reference can be applied to theories of abstract concepts in general (Dove, 2009), but also that we can use semantic externalism to avoid the conceptual mismatch problem.

However, if social kinds are real and mind-independent in the relevant sense, there is no reason to think that we cannot, at least in principle, track the respective social kind in the same way we track chemical kinds like water. We can track social kinds by means of their correlated properties, which are produced by (often not yet fully understood) social mechanisms that underlie the respective kind. Moreover, if social constructionists of *race* and *gender* are right,

superficial properties like gender performance and clothes may be even more directly and closely related to the corresponding kind than is the case for biological natural kinds like *water* (Löhr, 2019). While *transparent liquid* is only contingently connected to H2O, performance and clothes, i.e., gender expression, may be more essentially part of certain social constructs, such as gender.

Thus, at least in principle, externalist theories of concept acquisition can be applied to social constructs for which the mismatch problem is most pressing, namely covert social constructs, like those pertaining to gender, sexuality and race. We acquire the concepts of such supposedly covert kinds (i.e., the ability to think about them) by tracking and baptizing an observed regularity. Because this regularity can be explained by a common mechanism, it can lead us to the property in question just like *transparent liquid* can lead us to *water*. This regularity, in the case of social kinds, is a cluster of perceptually accessible properties that frequently co-occur. The co-occurrence of the properties that cluster together is, according to social constructionists, to be explained by a social mechanism, such as a set of beliefs commonly held in the community, certain traditions or institutions (Guala, 2016). In many cases, however, this mechanism is still to be spelled out in detail or discovered by future research in the social sciences.

6 The qua problem

This leaves us the with the qua problem for covert social kinds. I argue that the qua problem has a surprisingly simple solution once we assume that covert social kinds are stable and mind-independent kinds and we allow some mental representations that enable us to track this kind. Since, on the present account, we do not merely rely on an initial baptism to acquire the respective concept, but on a number of superficial properties that reliably co-occur (under the assumption that this co-occurrence is caused by a common mechanism), there are not many kinds in the vicinity that our concepts could refer to. For example, what makes it the case that I refer to men when referring to a certain regularity (certain co-occurring behaviors and clothing for instance) and not similar (but perhaps also natural or scientific) properties like *human* or *biped* is that none of these other properties explain the observed regularity.

This is again analogous to the concept of water. I can refer to *water*, as opposed to *liquid* because the perceptually accessible properties I identify as salient (e.g., that water is a liquid if above 0 degree Celsius, and frozen below 0 degree Celsius etc.) are explained by a common mechanism, in this case the common molecular structure H2O. Since *liquid* is not the natural

kind that explains these regularities (other liquids have been observed to have different properties), WATER refers to *water* and not to *liquid*. I thus cannot decide about what my natural kind concepts refer to. I might have the wrong belief that my prototype of water refers to XYZ and yet have the concept of water (as H2O) simply because my prototype of water is not explained by XYZ but H2O. Similarly, my prototype of a men leads me to the concept of man (as opposed to some other non-inductive or non-natural property) because there is a reason why men tend to have certain properties (namely certain social mechanisms).

Finally, the reason why we are able to detect these regularities in the first place and also why we are able to look beyond them (i.e., why we do not simply have a concept of prototypical men, but of men) is explained by several cognitive biases that have been found not only for concrete biological kinds, but also for social kinds. In particular there is convincing evidence for psychological essentialism not just for biological but also for social kinds like *race* (Gelman, 2003), which can be used to explain biased behavior towards members of an outgroup. So, the reason why we acquire the concept of white person, and not of prototypical white person (based on certain regularities) is that people are born with the bias that there is some underlying biological essence that explains these regularities.

In addition, core systems of agent representation (Spelke et al., 2013), joint attention and an innate motivation to cooperate (Tomasello, 2013) may explain how children have such little difficulty acquiring even highly abstract social kind terms at a very young age (Tardif et al., 2008 find the word 'naughty' to be among the first ten words produced by children) even if the beliefs associated with these terms may still be limited or even false. These innate social capacities play a similar role for the early acquisition of social kind concepts (similar to the whole object bias, for instance, in the case of functional kinds like *ball*): they allow the child to acquire a sense of which social properties are relevant and which can be ignored. This then allows for the quick acquisition of relevant world knowledge and regularities that suffice to meet the possession conditions of the respective (often highly abstract) social kind concept, including concepts like MAN or WHITE PERSON.

So, realism of covert social kinds, combined with plausible psychological social biases, make social constructionism compatible with externalist theories of reference. Such accounts explain how we get from superficial properties to their respective kinds without running into the conceptual mismatch problem or the qua problem. The former is avoided by the external link to a real mind-independent (in the relevant sense) kind. The latter is avoided by the kind itself

combined with some basic internal representations and biases. In other words, we can refer to women (and not to individuals similar to woman) because we are predisposed to detect certain regularities in our environment and based on other previous knowledge that allow us to establish the right causal relation to the social mechanisms that explain these regularities.

7 Conclusion

I argued that in order to apply semantic externalism to the conceptual and the linguistic mismatch problem faced by social constructionists, covert kinds (i.e., the kinds for which these problems are most pressing) must be stable and mind-independent in a rather strong sense. This excludes the possibility of a Searlian account of these kinds, according to which beliefs consisting of concepts of these kinds bring these kinds into existence. I argued that the strong realist and other psychological demands can be met by semantic externalism about concept acquisition in combination with a homeostatic property cluster view of social constructs. Both provide an adequate response to the mismatch problem and a plausible and attractive account of the acquisition of concepts of socially constructed kinds.

Conclusion

1 Main claims

One of the main claims of this thesis was that abstract concepts, e.g., concepts of social constructs, are not as special as we might think. In particular, I argued that they do not pose any *conceptual* threat to situated theories of cognition, e.g., empiricist or situated theories of concept application. I argued in chapter 3 that what makes the application of situated accounts to abstract concepts seem implausible is a conceptual conflation of that which determines the content of a concept and that which we use to *actually* (as opposed to *merely correctly*) apply the concept (in categorization or analogy-making, for example). The same distinction was applied to make sense of an externalist theory of concept acquisition, especially of concepts of social constructs in chapter 4.

In the first chapter, I argued for a distinction between what I called "concepts", i.e., what we think in terms of, and what I called "categorization devices", i.e., that which we rely on when applying our concepts in thought to objects. I argued that if we make such a distinction, we can see a number of issues much more clearly that have plagued the concept literature for decades. Most importantly, I argued that we can see why a theory of categorization devices need not give us an account of the strong semantic compositionality and systematicity that many philosophers assume necessary for a successful theory of concepts. This means that we can endorse psychological accounts of "concepts" as valuable contributions to theorizing about categorization without thereby expecting a full-fledged theory of concepts.

The distinction between concepts and categorization devices should not be confused with the claim that no concept can be individuated by means of its application or possession conditions and it should also not be confused with the claim that categorization devices are some sort of second content of a concept. Instead, a categorization device simply stores that which we use to actually decide whether a concept does or does not apply, whether this application is correct or not. It does not *determine* whether a concept application is correct or not. Correctness is either determined by definition if we have to do with a complex concept or the world-mind relation if we have to do with a simple (not composed) atomistic concept.

In the second chapter, I argued that there is no empirical or conceptual reason to think about categorization devices, which, in this paper I called "concepts" (following the psychological use), as stable entities. This claim was defended especially by Edouard Machery (2009) based

on the wrong assumption that categorization devices need to be stable to explain communication and stable behavior. Having the distinction from chapter 1 in mind, this worry completely dissolves. Moreover, I showed that the alternative of context-dependent sets of beliefs makes the same empirical predictions as the best defense of (categorization) invariantism, while invariantism, especially if it wants to explain abstract concepts, collapses into (categorization) contextualism.

In chapter 3, I discussed in detail how we can make sense of the idea that situated theories of concepts can explain abstract concepts, again based on the distinction introduced in chapter 1. In particular, I distinguished three kinds of objections to situated cognition from abstract concepts, i.e., three scope objections. I argued that once we distinguish the objection that situated theories cannot explain concept application and acquisition from the objection that they cannot explain what determines the content of our concepts, we can see that situated theories of cognition, as empirical theories in psychology, as opposed to conceptual theories in meta-semantics, do not face any particular conceptual problem explaining how we apply abstract concepts. They can still be empirically false, however.

In chapter 4, I applied the ideas from the previous chapters to the issue of concept acquisition of abstract concepts. I argued that we can explain the acquisition of certain abstract concepts, namely concepts of social constructs, by means of situations and experiences with concrete instances of the category. I showed that such an application is possible and applied externalist views of concept acquisition to socially constructed kinds. This chapter also included the argument that if we want to explain the acquisition of many abstract concepts in terms of an externalist strategy we need to assume a very strong realism of many abstract kinds, especially social kinds like *race* and *gender*.

Now we can put all the pieces together. Concepts are that which we apply to objects or that we relate to other concepts *in thought*. They are individuated by means of their content (reference) and their syntactic form. Concepts can either be simple or complex. They can be either descriptive if they are complex or non-descriptive if they are simple. Both claims are relatively uncontroversial. What is controversial is how many lexical concepts are of which kind. I have not taken a stance on this issue here and consider it an empirical question. Still, I take it to be empirically rather clear that while most concepts are complex, most lexicalized concepts are either simple or its associated words are polysemous. This, I assume, simply follows from the observation that we cannot define most ordinary language terms.

Concepts are applied by means of sets of beliefs or "bodies of knowledge" that I call "categorization devices" or "abduction devices", which consist at least partly of concepts and our attitudes toward them. There is no reason why we have to assume that what these devices store has to be context-independently retrieved and stored as a stable set. There is also no conceptual reason why the beliefs stored in this device cannot be grounded in perceptual areas as argued by neo-empiricists. What led us to such assumptions is a conflation of concepts with their categorization devices.

One main question that concerns us in the philosophy of psychology of concepts is how our concepts get their content. Again, I argued that there is no reason we cannot apply the simulation theory to explain how we apply abstract concepts even if it cannot explain how the concepts applied get their content. I argued that to possess many abstract concepts all we need is to establish a relation to a real kind or possess the right descriptions. So, again, concepts are either simple or complex entities and that which we use to apply concepts are context-dependent simulators, i.e., sets of beliefs or other intentional states representing rather concrete properties and situations at least partly in a modal format.

There is one question that remains to be addressed and that I am often asked about when presenting my distinction between concepts and categorization devices. The question is how exactly do concepts and categorization devices relate to each other? In other words, how does that which we think in terms of and that which we use to apply that which we think in terms of relate?

This question is surprisingly complex and has so far not been addressed in the literature. It is, however, not a question that is only relevant if one subscribes to the distinction between concepts and categorization devices. Of course, 'concept' and 'categorization device' are merely labels for independently interesting notions, i.e., that which we apply in thought and that which we use to make such applications. So, whether or not the reader agrees that philosophers are mainly interested in the former and psychologists are mainly interested in the latter, the distinction described in this thesis is of independent importance even if one assumes that both descriptions usually pick out the same entity (i.e., even if one disagrees with one of the main claims of this thesis).

2 How do concepts and categorization devices relate?

How does the notion of categorization device relate to the notion of concept? The short answer is that categorization devices (that which we use to apply our concepts), if they are sets of beliefs (they might include other kinds of representations), are at least partly made up of concepts. If, as I have assumed throughout the thesis, entertaining a belief is simply applying a concept to an object or to another concept (with a certain attitude), then, in a metaphysically innocent sense, categorization devices, if they consist partly of beliefs, partly "consist" of concepts. A proper answer to this question however runs deeper. It is best answered by comparing how both concepts and categorization devices are individuated. This question was already addressed in chapter 1, but with a focus on the differences of concepts and categorization devices. I now would like to focus on giving a more positive account of their relation to each other.

Simple concepts are individuated, at least partly, by their reference.[34] The concept of dog, if it is simple, is essentially about dogs. If it were about typical dogs it would be the complex concept TYPICAL DOG. Similarly, the concept of financial institute and the concept of riverbank may share a single lexical form 'bank' but they are clearly different as they are about fundamentally different kinds of things. However, sometimes the same kind of thing may present itself in very different ways suggesting that reference is not enough to individuate simple concepts. This can be illustrated especially well by considering so-called "Frege-cases".

Frege-cases are thought experiments where the same reference (e.g., Superman or Angela Merkel) is presented in different contexts in fundamentally different ways to the subject (e.g., one time as superman and another time as Clark Kent or one time as loving wife and another time as the chancellor of Germany) without the subject knowing that she has to do with the same person.

Frege-cases have convinced most philosophers (including Fodor, 1998) that we cannot individuate concept-types (not to be confused with "types of concepts") as well as belief-types by means of their reference (what individual, extension, property or relation the respective representation is about) *alone*. The main reason for this is that perfectly rational people can hold

[34] Descriptive concepts, i.e., complex definitional concepts, such as THE FIRST MAN ON THE MOON, are individuated not by their reference (the reference may change in different possible worlds). Instead, they are *essentially* individuated by their form alone. Two different complex concepts are essentially different concepts. Whether they pick out the same things is irrelevant.

conflicting beliefs about the same thing without knowing it, *even after reflection*. The reason this is possible is that we are often presented with the same thing from different perspectives, i.e., the same object can have different so-called "modes of presentation", i.e., ways it is being presented to us.

For example, we may know a lot about Toni Morrison, the nice neighbor next door, and we may know a lot about Toni Morrison the author. However, since we know that 'Toni Morrison' is a common name and we know that more than one person can share the same name, we will, if we are rational, not believe, *even after reflection*, that both names denote the same individual. So, it is perfectly rational to believe that Toni Morrison, the neighbor, is probably not the best-selling author even though, in reality, they are the same person.

Most philosophers assume that Frege-cases show that I have two distinct beliefs-types with different contents when I say that "Morrison is a great author" is false with respect to my neighbor and true with respect to the author. Most contemporary philosophers of language and mind also agree (mostly due to reasons of compositionality, see Fodor, 1989) that what makes it the case that the different belief-tokens expressed by "Toni Morrison is a great author" are in fact of different belief-types is that they consist of different concept-types. In other words, it is assumed that I have two concept-types of Toni Morrison each of which is conjoined with a different set or body of beliefs. What explains the difference in my behavior towards Morrison, the neighbor, and Morrison, the author, is that I have different beliefs associated with the concept of each that are not linked (in certain interesting ways relevant for assessing the rationality of the subject) even after I think long and hard whether both might denote the same individual.

It is important to emphasize that the postulation of two concepts of Toni Morrison is neither a mere philosopher's stipulation, nor based on thought experiments alone. Frege-cases reflect an empirical fact about our psychology, namely about how knowledge is organized in the mind. They capture the empirical and psychologically relevant fact that some beliefs are not mutually available even after deliberation and even though they are about the same object. This phenomenon is of course extremely widespread and goes far beyond typical Frege-cases. I currently believe, for example, that gold and water are fundamentally different things and that what I believe about water is not true of gold. I could be wrong and my concept of gold and my concept of water are co-referential. However, I currently have no reason to make this assumption as both present themselves in fundamentally different ways. It would be irrational

to do so without justification. This observation must be reflected in a theory of how my knowledge about the world is organized in my mind. Concepts and thoughts must be individuated by more than just their reference.

Thus, Frege-cases, besides the various philosophical questions they raise, basically pose a question of how our beliefs are organized in the mind. The way knowledge is organized in the mind is of course not only what philosophers but in particular what cognitive psychologists are interested in. Frege-cases are thus not merely a conceptual issue in philosophy of language, but important devices for psychology. Frege-cases show us that there can be something like a "weak informational incapsulation" (my terminology) even when we have to do with the same object in the world. By weak informational incapsulation I simply mean that some information is not available when we think about some object even after deliberation.

I argue that we can model this weak-encapsulation of information by saying that we can have two simple concepts (in the case of individuals and natural kinds at least) or complex concepts (in many other cases) with the same reference and that each of these concepts has a different categorization device (they are applied by means of different sets of beliefs). Each categorization device is informationally weakly encapsulated from the other one, meaning that it would be irrational to merge them (unless we had good reasons to do so, e.g., if we learned that my neighbor is a best-selling author we might finally make the connection). Thus, our concept of Toni Morrison, the writer, is associated with the categorization device that stores the belief that Toni is a great writer, while our concept of Toni Morrison, the neighbor, is associated with the categorization device that stores the belief that Toni is a quiet neighbor. Categorization devices are thus simply individuated by means of the concepts they apply as opposed to by means of the beliefs or information they store. In other words, every concept has a categorization device, even if it is empty, and every categorization device applies a single simple or complex concept.

Concepts (as constituents of thoughts) are thus important to individuate categorization devices and Frege-cases teach us that each concept has at least one categorization device, even if it is empty (if it does not store any intentional states). Of course, there might be other constraints on how to individuate categorization devices discovered by empirical research. According categorization contextualists like Barsalou (1999), the empirical evidence suggests that we need to divide categorization devices into "simulators" and temporary, i.e., context-dependent "simulations" based on such simulators. According to Machery (2009; 2015), the same

empirical research (see chapter 2 for a discussion) in psychology strongly suggests that we need to split that which explains e.g., categorization, i.e., our categorization device, in two, namely in default-categorization device and background knowledge. The former is retrieved by default, constituents of the latter are only available in some contexts.

So, default-concepts *a la* Machery cannot *just* be individuated by means of which concepts they apply (see again chapter 2). In order to individuate a "default-concept" (default-chategorization device), according to Machery, we also need to identify the subset of our world knowledge that is retrieved by default. Machery (2009) argues that we can find out which structures are in fact default structures used in higher cognition by means of the following two heuristics: First, we have to do with at least two "default-concepts" if retrieving the one set of beliefs does not automatically and immediately retrieve the other set. A second heuristic is that if people make contradicting statements, such as "tomatoes are vegetables" in one context and "tomatoes are fruit" in another, this, so Machery, is evidence for two distinct default knowledge structures or default-categorization device. Note that these are mere heuristics and not necessary or sufficient conditions.

Machery's heuristics sound like versions of Frege-cases. However, Machery is not talking here about how to individuate thought contents or concepts (see chapter 1). He is not even giving us necessary and sufficient conditions of individuation. Instead, he is merely giving us some instrumental principle that can help us to empirically find the structures that allow us to apply certain concept-tokens, i.e., how we can locate default-categorization devices. So, if two sets of knowledge associated with the same concept-token are not immediately and simultaneously available, then we probably have to do with two distinct default- categorization devices. However, this does not mean that *even after deliberation* we do know that both sets are actually associated with the same concept token. For example, most people who know that tomatoes are in fact fruit can easily retrieve the information that, in a way, they are also vegetables. Machery's main example of how to find out whether we have to do with distinct default-categorization device is not a Frege-case.

Importantly, the notion of categorization device, i.e., including the notion of default-categorization device and including the simulator/simulation distinction Barsalou (1999) favors, is *not* individuated by means of the particular beliefs they store. This means that the default-categorization device of dog does not suddenly turn into a different categorization device of dog once we add or subtract a belief, such as, that dogs smell or that dogs are furry.

Similarly, the Barsalou-simulator of dog remains the same simulator even after an exposure to a new atypical kind of dog that may be represented as an exemplar representation. In other words, the categorization device (whether default-categorization device or simulator) of dog is *essentially* the categorization device of dog because it is used to apply the concept token of dog. **"**Essentially" here means that if it were the default-categorization device of cat it would be used to apply the concept token of cat. Categorization devices are not individuated by means of the beliefs they store.

To individuate categorization devices (including default/background-categorization devices and simulators/simulations) by means of which concepts they apply and not by means of which beliefs they store has two important advantages, one empirical and one methodological. Most importantly, to individuate categorization devices by means of the concepts they apply and not by means of the beliefs they store allows for the same set of beliefs to apply different concepts. I might hear that Aristotle is a philosopher and teacher of Alexander the Great on one occasion and that Aristotle is a philosopher and teacher of Alexander the Great on a different occasion and still be able to rationally ask whether we have to do with the same person. So, we need to posit two categorization devices because the merging of both devices would still be informative. This was not the case for the tomato case where we already knew that tomatoes are technically fruit even though this knowledge is not usually immediately available.

A second advantage of individuating categorization devices, i.e., default/background-categorization devices, simulators/simulations and so forth, by means of their respective concepts and not by means of the information they store is that it allows us to make sense of inter-personal and even intra-personal differences while remaining terminologically and ontologically parsimonious. For example, while a sighted person might apply their concept of dog only by sight, a blind person might apply the same concept type of dog only by means of auditory information. Were we to individuate categorization devices by means of the beliefs they store, we would have to conclude that both individuals have different categorization device types associated with the same concept type DOG. However, the justification for this way of modeling long-term memory is not clear and far less elegant considering that all we want to say is that the beliefs both individuals use to apply the same concept type are different. To say this, we can simply say that the intentional states they store in their respective categorization devices of the same type are different.

Note, however, that while for Machery' default-categorization device, there can be inter-subject variability, there cannot be, by definition, any within-subject variation across a short amount of time (again at least for Machery). Conceptual change, in the sense of changing sets of beliefs, is therefore only acceptable by Machery as a long-term change or as the effect of intense learning. For Machery, in order to be a default-categorization device, the set of beliefs that are stored in a default-categorization device must be stable within subjects. This is not the case for context-dependent Barsalou-simulations and simulators. It is also not the case for the notion of categorization devices, which leaves open to empirical research whether or not the set of beliefs stored are stable within subjects. However, again, I take it that, based on what we know from psychology, categorization devices tend to be context-dependent at least to some extent (see chapter 2).

3 Language and concepts

There is a lot missing in this thesis for a full theory of abstract concepts. In particular, what is missing is a theory of abstract language. Thus, most of my current work and my post-doc project focus on language. I argue that distinguishing semantic questions of content from epistemic or psychological questions of understanding and application allows us to make progress in studying how we are able to think and especially talk about very abstract things like love, gender or philosophy.

I argue that with a distinction between concepts and categorization devices, the explanation of linguistic behavior (in particular how we can bring the psychology and the philosophy of language together) can be rather simple: people are able to use and understand linguistic expressions in rule governed ways. Linguistic expressions are associated with concepts as well as a number of beliefs of correct applications, norms, feelings, episodic memories and so forth. The question of how we use linguistic expressions should be separated from questions of the meaning or content of these expressions and we may use a word without knowing much about what it correctly applies to, or without knowing the description that it *is correctly* associated with. Still, prototypes and exemplar or simulations, as I argue, may suffice to explain how we are able to use these words appropriately to coordinate our behavior with others.

Take for example the sentence "Germany is not a democracy". This sentence is a paradigmatic abstract sentence that is difficult to really explicate by many of those who might utter it. Just ask anyone who is not a political philosopher what this sentence could precisely mean, that is, which thought it might express. In psychology, sentences like this one are rarely studied.

However, more concrete sentences like “birds can fly” or “dogs bark” are studied extensively. Both linguists and psychologists usually assume that ordinary people understand these sentences by conjoining meanings or concepts together to construct the meaning of sentences. While this sounds relatively unproblematic for many concrete sentences, it is difficult to apply to the more abstract sentences. It is not clear whether ordinary but proficient speakers represent a meaning for ‘Germany’ or ‘democracy’ especially if, when asked, most people will probably merely respond with a loose set of properties they associate with Germany and what they take a typical democracy to be like. This does not mean, however, that the combination of these properties and conditions of a typical democracy are what the sentence means or even what the speakers mean when uttering the sentence. Again, we should distinguish that which allows us to apply an expression from its meaning (we should distinguish categorization or application devices from concepts).

What the meaning of “Germany is not a democracy” is, is controversial. I take it that the term ‘democracy’ is polysemous. What the sentence means in a given context needs to be negotiated and cannot simply be considered the output of ordinary people’s linguistic faculty. To understand the sentence “Germany is not a democracy” to a degree that allows us to respond to it appropriately in context however is much easier to determine. In a context for example in which the sentence is uttered by a friend who is frustrated that the German government does little to prevent rents from rising in its major cities, little agreement on the meaning of terms is required for a reasonable response that represents the speaker’s own political stance. All that is required is to associate the term ‘Germany’ with the belief that it denotes the place one is currently living in and ‘democracy’ with the stereotypes of a democracy, say, free speech and free elections. The speaker can then respond that Germany is still a democracy because in her episodic memory she has many experiences of free elections and free speech. The other speaker may then respond that she disagrees and that the people who vote for the conservative dominant parties are to blame for the overly capitalist system that the interlocutor complains about.

Thus, the idea is that breaking down the understanding of abstract sentences to concrete experiences can allow us to de-mystify abstract language and thought and the separation of content and application can avoid traditional objections to experience-based accounts of communication. What content of our sentence we are eventually committed to is a different and more difficult question that may not be decided before we can engage in successful linguistic behavior. This is the case because especially abstract lexical expressions are often associated with a number of different simple and complex concepts. In other cases, it may be that people

use words, especially abstract words, without there being any associated concept. That this is the case is especially clear for words like “Yay” or even “hello”. Here it is not clear whether we have to do with any thought that is being expressed. The same could be argued for such difficult words like “irony” or “anxiety” that we might use in proficient ways without it being clear whether the proficient user really associates a determinate concept with these terms.

To give a concrete example of how the distinction between content and categorization can be made fruitful for the empirical study of language, I would like to propose in the next section, an application of this view to language. I argue that the distinction allows us to use even a simple simulation theory – that could in principle be grounded in perceptual and emotional systems – to explain our use of abstract language in a phenomenon called “copredication”. What a copredication sentence means, I argue, may be a different question that can be tackled in a post-hoc manner after the sentence was understood and proficiently used.

4 One linguistic application: simulations and abstract copredication

4.1 What is copredication?

The term 'copredication' is commonly used to capture the phenomenon that we can, arguably, use a single expression to denote two distinct but related entities in the same sentence with different predications. To illustrate, compare the following sentences:

(1) John entered and left the bank.

(2) The bank is large and unfriendly.

(3) John entered and sold the bank.

In the case of (1), a single noun 'bank' is copredicated by 'entered' and 'left'. Since both predicates arguably denote the same entity, the building of the bank, this example is usually considered relatively unproblematic.

In the case of (2), the single noun 'bank' is again the argument of two different predicates, 'large' and 'unfriendly'. However, unlike in the previous case, it seems that the two predicates denote two different entities associated with 'bank', namely the building in the case of 'large' and the staff of the bank in the case of 'unfriendly'. Since both objects are, arguably, concrete physical objects, I call this kind of copredication “concrete-concrete copredication”.

In the case of (3), the predicate 'entered' applies to a physical object, the building, while the predicate 'sold', at least in this context, is meant to apply to the more abstract financial institution that is rather independent of the concrete building that hosts it. Since the first object is a concrete physical entity and the second an abstract kind, I call (3) an instance of "concrete-abstract copredication".

Copredication is usually analyzed in terms of formal models of polysemy, but I would like to argue that simulation theory can give a much simpler account at least of linguistic understanding, especially of concrete-abstract copredication. Applied to copredication, the simulation theory of linguistic understanding predicts that the more difficult it is to integrate the representations that become readily available upon being exposed to a copredication sentence into a coherent simulation of a concrete situation, the more this sentence sounds anomalous.

The primary question raised by copredication is why some copredication sentences sound felicitous to us even though the relevant predicates apply not only to different entities as in (2), but even to ontologically fundamentally different kinds, as in (3). This is the primary question that a theory of copredication has to answer because if (2) and (3) did not sound felicitous to us, copredication would not raise any especially interesting metaphysical or semantic issues.

Second, a theory of copredication should be able to explain why some copredication sentences that appear very similar to (3) do not sound felicitous. Consider the following examples:

(4) a. The newspaper that fired its best journalist fell off the table.

b. #The newspaper fired its best journalist and fell off the table.

(5) a. Anna opened, read and sued the newspaper.

b. #Anna opened and sued the newspaper.

(6) a. The bank was set on fire after hiring a new sales executive.

b. #The bank was set on fire after flying to the Cayman Islands.

Why do we find (4a), (5a), (6a) and (7a) acceptable, but (4b), (5b) and (6b) anomalous?

Third, a successful theory of copredication should explain not just why and when copredication sentences sound felicitous, but why we immediately understand them *in a certain way*. Take the following examples:

(7) a. The rabbit was killed, eaten and worn.

b. Berlin is big, dirty and never sleeps.

There is no metaphysical reason why we could not understand (7a) as meaning that the rabbit was killed, eaten and worn as a whole. In a different context, this is exactly what we would expect. Just imagine a monster killing and eating rabbit as a whole. Similarly, why do we understand (7b) immediately as saying that Berlin's size or population is big, that its streets are dirty and that some of its bars have long opening hours if nothing in the sentence explicitly suggests this interpretation?

Furthermore, a successful theory of copredication has to explain why we individuate contents intuitively and immediately *in a certain way*. Consider the following sentence:

(8) John picked up and mastered three books.

In (8), the books in question are, intuitively, neither individuated in terms of just a physical copy (John probably did not pick up three physical copies of the same book and mastered each one after the other), nor do we individuate 'book' just informationally (because we usually do not assume that John picked up a trilogy). Instead what the speaker probably means is that John picked up and mastered three both informationally and physically individuated books (cp. Gotham, 2014). But why does this interpretation strike us as the most plausible?

Fourth, copredication poses a challenge to a theory of linguistic processing, in particular how we construct sentence meanings. According to the traditional picture of linguistic processing and one that is still often assumed in formal semantics and philosophy of language, every expression contributes a single meaning to the meaning of the sentence. In the case of copredication, however, it seems that a single expression can contribute several senses to the meaning of the sentence, some of which only become relevant in certain contexts.

Finally, copredication poses a challenge to a straightforward and traditional theory of meaning as involving external objects in the world. Most prominently Chomsky (e.g., 2000), but also Collins (2009) or Pietrosky (2018), argued that such an externalist theory of meaning would

commit us to the existence of such strange things as banks that can both be entered and sued or entities that can hold a liquid and be the liquid itself (as in "Tim drank and dropped the bottle"). Such objects, these skeptics argue, do not exist, but still the sentence is not meaningless. Consequently, the meaning of a word cannot be its extension.

4.2 Simulations and situation models

I argue that the simulation theory in combination with an independent theory of content, i.e., presupposing the concept-categorization device distinction, can meet all the above desiderata, including Chomsky's challenge, and thereby offering a genuine and novel application of the theory proposed in this thesis to more traditional problems in linguistics and philosophy of language. In the next two subsections, I will briefly propose a way to account for the psychological desiderata above with a simulation theory. In the final subsection, I give a separate proposal as to how we can answer problems for a theory of content raised by copredication. However, I wish to note that the simulation theory developed here can be combined with any theory of content recently proposed in the literature on copredication (e.g., Gotham, 2014; Liebesman & Magidor, 2017; Asher, 2011).

Again, simulation theory predicts that sentences are more felicitious if we can easily generate simulations of concrete situations based on these sentences and the context. The question of what determines the ease of integration of a new representation into a simulation of a concrete situation is a complex empirical matter that is not well understood (Zwaan, 2016). However, at least in the case of copredication, two plausible variables or dimensions that make insightful predictions can be derived from the situation model literature (e.g., Zwaan et al., 1995). Following the situation-model literature, I call the first "spatial contiguity" and the second "causal contiguity".

Spatial contiguity, in the situation-model literature, is usually understood as a more global measure that concerns changes between rooms or other larger areas that are perceived as dis-contiguous by the viewer. For example, it is assumed that individuals experience spatial dis-contiguity if a narration features a rapid jump in locations, which surfaces in terms of delay in processing and understanding (e.g., Ehrlich & Johnson-Laird, 1982).

We can use the notion of spatial contiguity for more local, sentence-based phenomena. Consider the following two sentences. The reason, according to the present approach, we can easily understand (9a) but not (9b)

(9) a. The glass door was opened and John went through it.

b. #The glass door was being repaired and John went through it.

is that in (9a) we are "instructed" to simulate a concrete glass door opening and a person walking through the opening revealed by the movement of the door. This is compatible with our readily available prototypical world knowledge concerning doors and human movements. We find it easy to simulate a concrete coherent situation even though (9a) does not make the connection between the two predications explicit.

Example (9b) probably generates a simulation of a door in a horizontal position considering that we assume that doors are typically repaired horizontally, while it generates a representation of a person in a vertical position. This dis-contiguity of orientation makes it difficult to connect the two sub-simulations into a coherent simulation of a possible scenario. Simulation theory predicts that since it is not easy to think of a situation where a person can walk through a horizontal door, i.e., since there is no spatial cohesion between the representation of the door and the representation of the event of walking through the same door, (9b) sounds anomalous.

However, simulation theory also predicts that all that is needed to make (9b) sound felicitous is additional contextual information that would make the relation easier to simulate. For example, we could imagine a science-fiction context in which John can walk on and through walls. In such a context (9b) sounds arguably less anomalous because the connection between the first and second predication is easier to simulate, i.e., we can more easily imagine John walking through the horizontally oriented glass door. We could also simply add the information that the door is being repaired in a vertical position.

In the case of causal contiguity (Gernsbacher, 1995), simulation theory predicts that individuals experience causal dis-contiguity in narrations if a direct causal link between two events is not sufficiently established. There is ample evidence that causal dis-contiguity delays processing considerably (e.g., Magliano et al., 1993). Again, we can apply the same idea to more local, sentence-based phenomena. Consider, for example, the following sentences:

(5) a. Anna opened, read and sued the newspaper.

b. #Anna opened and sued the newspaper.

The reason why (5b), according to simulation theory, is anomalous in a neutral context (one that does not provide additional relevant information) is that it is difficult to make the causal link between Anna opening a physical paper and filing a lawsuit against the newspaper. Since it is difficult to generate coherent situation model consistent with (5b), even after some time processing, we tend to find this sentence anomalous.

However, again, simulation theory also predicts that surprisingly little is required to make (5b) sound much less anomalous. One strategy for repairing the sentence is to provide a missing causal link between both predicates as is done in (5a). Importantly, the difference between (5a) and (5b) is not that the former is *a lot* more explicit. (5a) is just explicit enough for the hearer to make the right causal connection. The hearer still has to "fill in the gaps" that the speaker expects to be taken for granted by both interlocutors. However, the speaker has now made it easy enough for the hearer to integrate both predicates in a coherent simulation while still avoiding redundancy.

The added verb can be omitted with the felicity of the sentence preserved if (5b) is presented in a discourse context that provides additional relevant information, e.g., that Anna is a controversial celebrity. This strategy for making the sentence much easier to simulate, and hence more felicitous, builds on the world knowledge that celebrities are often victims of smear campaigns in newspapers.

4.3 Abstract copredication

As I have discussed above in the introduction, one problem with the simulation view of copredication and simulation theories in general is that it is not immediately obvious how they can account for sentences that involve concepts and words that do not refer to concrete or highly imaginable entities, such as abstract entities like democracies, financial institutions or truth. Again, if the simulation theory can only explain concrete copredication sentences, its scope is extremely limited considering that the most interesting copredication cases involve both abstract and concrete objects.

In this section I develop the idea of abstract concepts in terms of the simulation of concrete situations further. Take for example again the expression 'newspaper' that can be used to denote both the physical paper as well as the company or institution. According to a simulationist approach to copredication, the reason we can say (4a) but not (4b)

(4a) The newspaper that fired its best journalist fell off the table.

(4b) #The newspaper fired its best journalist and fell off the table.

is, first, that in a context in which I am asked which newspaper I am supposed to hand over, the word 'that' in (4a) instructs me to use a certain part of my world knowledge specified in the that-clause (the firing of the journalist) to simulate a single specific object, namely the physical paper that is associated with this aspect of my world knowledge. Since we can easily simulate a single object falling off a table (4a) sounds felicitous (spatial and causal contiguity remain intact).

A more interesting question is why (4b) sounds anomalous. According to simulation theory, the reason (4b) sounds anomalous is that the beginning of the sentence does not instruct us to simulate a single concrete object that could easily be simulated as falling off a table. Instead, upon hearing the term 'newspaper', we likely retrieve a number of representations of concrete objects that fit to the first predicate. Concretely, hearing the term 'newspaper' in the context of 'firing a journalist', we likely simulate a concrete situation that involves human beings in offices with computers. Integrating a set of people in an office space with something that can fall off a table into a coherent plausible simulation is not easy (it displays causal and spacial dis-contiguity), hence the sentence sounds anomalous. I argue that this sentence sound anomalous is evidence (albeit not conclusive evidence) that we in fact understand abstract words in terms of concrete representations.

Similarly, consider the following assertion by one parent in a discussion about the best local schools:

(10) No school in the area offers classes in fine arts.

This sentence can easily be understood despite its highly abstract content consisting of a negation, the abstract institution sense of school, as well as the abstract concept of fine arts. The assertion invites the hearer to judge its truth value. Now, the seemingly abstract sentence (11) can be explained in surprisingly concrete terms. All the hearer now has to do to form an adequate response (expressing agreement or disagreement) is to search their episodic memory for whether they have ever heard or seen anything related to children painting in a classroom.

That we really understand and react to abstract expressions in terms of concrete situations is further suggested by the felicity of the following sentence:

(11) On September 11, the newspaper was completely shocked by what had happened.

Since, according to simulation theorists, 'newspaper' in some contexts makes available representations of people and spaces, it is easy to simulate a situation in which people that we conceptualize as journalists in an office space behave in ways that we conceptualize as being in shock, e.g., by imagining them crying.

Similarly, according to simulation theory, the reason we find (12a) but not (12b) felicitous

(12) a. The newspaper that fired its agency fell off the table.

b. #The book that fired its agency fell off the table.

is that 'book' is simulated as a mere physical object that cannot fire anybody (except in a fictional context), while 'newspaper' can be simulated in ways that represent a more abstract meaning understood, e.g., in terms of a group of individuals that *can* fire an agency. The question is again why we can observe such an asymmetry considering that, *prima facie*, books and newspaper are both the products of a single person or group of people.

One observation that might make this issue clearer is that there does not seem to be a single word denoting the more abstract sense of 'newspaper', i.e., the company. Hence, the practice of using this expression both for the physical paper and the company might have started as a metonymic modulation using the concept of a physical newspaper as a stand-in for the more abstract management or company that produces and distributes the newspapers. Since that which produces books are usually single authors and that which distributes books are usually publishing houses, for which we do have expressions, there was no need to use 'book' to lexicalize the more abstract sense.

In other words, the reason why there is an asymmetry between 'book' and 'newspaper' in terms of what strikes us as an acceptable sentence may be best explained in terms of their different levels of generality. One might assume that the speaker uses a word that is only specific enough to allow for a relevant and fast simulation. This predicts that if the speaker were to mean that the author of the book fired their agency, they would have just used 'author', while there is no such precise word available for 'newspaper'. Words like ‘management’, ’owner’, and ‘company’ are too general and 'management of the newspaper' seems to be overly specific and redundant.

This does not mean that, according to the simulation view, we do not store knowledge about how books are produced and who publishes them. However, this information is not and need not be accessed in this context (it is not retrieved "by default" to use terminology introduced by Machery, 2009). Instead, contextualist approaches to the retrieval and organization of knowledge (e.g., Barsalou, 1999 and see chapter 3) promise a more appropriate description of linguistic understanding of copredication.

Thus, copredication gives us support for the claim that abstract concepts and words can be understood in terms of very concrete representations, e.g., of simulations of concrete situations that are compatible with the truth of the respective sentence. Again, if true, this would not only give us a genuine explanation of the psychological questions raised by copredication, but also give us an intuitive account of how we can understand and use abstract words and concepts even though we do not have direct perceptual contact with their referents.

4.4 Understanding and content

Now the challenge for any simulation theory of linguistic understanding remains that the same concrete situation is compatible with a number of different sentence meanings and linguistic understandings. For example, even though the situation models that may explain our immediate behavior towards sentences (3) "John entered and sold the bank" and (3a) "John entered and bought the bank" might be highly similar (perhaps a person entering a building and shaking hands with another person in a suit), both sentences not only have different semantic contents, but we also *understand* them differently. Consequently, so a traditional worry, neither linguistic meaning nor linguistic understanding can be reduced to simple simulations of a single concrete situation (see the introduction for this worry).

I take these worries about the simulation view to be attacking both an overly ambitious and simplistic version of it. First, simulationism does not entail that we simulate a single static image for each sentence. Instead, we usually simulate a rich dynamic situation-model that includes movements and a number of alternative and possible future situations. The richer and more dynamic this model, the more is it able to distinguish the meaning of different sentences. For instance, while an understanding of the sentence "Two people are shaking hands" might involve simulating two people shaking hands, (3) will likely produce a much richer dynamic situation model, say of someone leaving the bank with money in their hands, celebrating with their friends and so forth. Upon hearing (3a), on the other hand, we might generate a model where a person remains inside the bank, giving instructions to their new employees.

Second, what explains the difference in understanding between (3) and (3a) is not necessarily just the simulations that are being generated upon hearing each sentence, but the distinct ways the respective hearer's world knowledge is updated. This updating may not immediately result in different simulations, but different resulting dispositions and consequently different future behavior and simulations. While (3) updates the hearer's world knowledge based on the representations made more available upon hearing 'sold', (3a) updates their world knowledge based on the representations that become more available after hearing 'bought'. From this updated world knowledge, which influences our expectations, we can then predict different future simulations and actions even though the initial simulations generated may remain less rich or even highly similar.

Most importantly, as argued in chapter 1, a theory of linguistic understanding in terms of simulations does not necessarily have to provide a theory of semantic content. In other words, we may argue that someone will understand and react to a sentence by means of many rich dynamic simulations that go beyond merely simulating people shaking hands, but simulation theorists do *not* need to argue that even this highly rich but subjective and context dependent simulation constitutes the semantic content or meaning of (3). After all, the sentence is about John entering and selling the bank and not about John shaking hands. Moreover, we would not say that when I think of a situation in which John sells the bank online while you simulate a handshake that I grasp a different meaning of (3) than you. Simulation theorists can agree that this would make communication very difficult if not impossible.

We can avoid these problems by distinguishing "linguistic understanding" (an epistemic notion denoting the integration of the representations made available by the sentence into a coherent situation model) from what I call "linguistic comprehension" (grasping the semantic content of a sentence). This distinction between epistemic states and semantic linguistic content should be familiar from other debates that make use of a strict distinction between epistemology and semantics (e.g., Putnam, 1975; Fodor, 1998). According to this tradition, we can grasp the meaning of a concept or word, i.e., "comprehend" the meaning of a symbol, without there being much understanding. Again, the same distinction applied to concepts was defended above.

To make sense of simulation theory with respect to questions of semantic content, simulation theory should construe situation-models as the starting point for determining the content of the expression, but not the endpoint. The idea is that, when analyzing the meaning of a sentence (which itself is a relatively rare and context-specific activity often confined to academic and

political occasions), we base our semantic intuitions on our simulations of possible situations that we take to be compatible with the respective sentence without yet having identified its meaning. Semantic content then becomes more a matter of negotiation rather than simply the output of our sub-personal compositional language module.

This account fits with the fact that the semantic content of a sentence is often very difficult to determine. What the meaning of a sentence is can be extremely controversial, as anyone who ever engaged in a philosophical debate about, say, whether the sentence "killing is bad" or "the mind is identical to the brain" is true, knows. To find out what a sentence means often requires negotiation rather than mere combination of associated concepts. So, I take it that even though simulations can explain our semantic intuitions and our linguistic behaviors, simulations may not give us the semantic content of a sentence and I do not take this to be the ambition of psychologists interested in simulations.

Thus, copredication may not be an exceptional phenomenon of language, but merely a particularly clear instance of superficial linguistic processing whereby the semantic content of the sentence need not be clearly represented by the speaker. Still, we can make sense of copredication sentences, i.e., understand them by means of generating models of situations that we take to be probable assuming the sentence is true based on the representations made available upon hearing the constituents of the sentence.

So, the reason we use 'bank' to refer both to the building and the management is not that 'bank' best represents what we mean. Instead, we choose 'bank' simply because we take it to be the most appropriate expression (not too specific and not too general) to easily elicit the right situation models in the hearer based on what we take them to already know from context and world knowledge. For example, we assume that 'bank' in combination with 'enter' will generate a model of a person entering a building because this is most compatible with our world knowledge. Hence, we do not assume that we have to specify that we are talking about the building of the bank, for largely neo-Gricean reasons.

The simulation view also gives an empirically plausible and intuitive explanation of interpretations of sentences like (7) or (8). The question here was why we interpret (7a) immediately as asserting that the rabbit was killed and its flesh eaten and why we individuate the book in (8) both informationally and by means of its physical instantiation. The answer is simply that world knowledge and context determine the most plausible model of the situation we think is being described by the sentence. Since we know that people usually eat the flesh of

a rabbit and we assume the sentence to be about humans we do not simulate a monster eating a whole rabbit but a person eating its flesh.

In the case of 'book', we assume that John picked up three different both informationally and physically individuated books simply because this is the most likely scenario based on the information we have. However, nothing in the sentence excludes other possibilities. The sentence could also be used to describe a situation in which John picked up a trilogy, even if this is not the interpretation that first comes to mind. Still, when analyzing the meaning of the sentence we usually begin with these prima facie more plausible situation-models, which then leads many linguists to try to construct a semantic formalization that captures these initial semantic intuitions.

Finally, the present account explains why we can state that (3) literally means that John entered the building of the bank and then sold the bank without having to commit to mysterious entities that can both be entered and abstractly sold thereby addressing Chomsky's traditional objection to truth conditional externalist semantics. We can simply say that the surface level of a sentence does not reflect the intuitive or minimal meaning of the sentence, which can, in a post-hoc manner, be negotiated in traditional truth conditional ways.

5 Final words

In this conclusion, I addressed mainly two remaining questions. First, how do concepts and categorization devices relate. I argued that in so far as categorization devices store beliefs or other propositional attitudes, they "consist" of concepts. Moreover, every concept has a categorization device even if it may not always store beliefs or other kinds of representations, i.e., if it is "empty". This, I argued, promises a new way to account for Frege's problem that does not require two kinds of conceptual content, say a "sense" and a reference.

Secondly, I proposed how we could apply this theory to problems in the psychology of language and linguistics. I argued that we can employ simulation theory to explain the phenomenon of copredication. However, the simulation theory has a major problem that has been discussed in the introduction and chapter 3: it cannot account for the content of our sentences and words. Here, my distinction between the content of concepts and the beliefs or simulations we use to apply concepts can help if applied to words. I argued that we may use our words based on simulations, which however does not yet settle their meanings. What the meaning of a sentence is, often needs to be negotiated in context and cannot simply be derived from its surface

structure. In other words, what concepts our sentences express is often a controversial matter. With the presently defended distinction between concepts, their content and the representations we use to apply our words and concepts, we can now make sense of this fact.

For the practice in philosophy the application of this theory is this: it may not be the case that we can infer the content or meaning of our ordinary words by reflecting on how we would apply them to paradigmatic situations or by trying to search for that which we use to apply this word. That which we use to apply a word will be guided by all kinds of world knowledge, context dependent representations and both moral and conversational norms. Moreover, the same word is often used in such different ways that we must assume that it is associated with a number of different concepts that may or may not be related. Thus, we should not assume that our linguistic expressions neatly line up with the same or approximately the same number of basic concepts. Instead, we can expect that it may be a matter of pragmatic or methodological choice how to divide the different meanings of a single linguistic expression into one or several concepts that may not match up with the content of categorization devices of these words.

Still, my hope is that an investigation into the psychology of how we *in fact* apply, e.g., the word and concept of knowledge and abstract concepts in general, can make some progress in the search for a method to do philosophy. This approach is heavily inspired by Wittgenstein's middle period, which I still take to include a simple truism: if we want to know what 'knowledge' or 'truth' mean, we should first look at how these terms are used. This approach however does not commit me or Wittgenstein to the metaphysical view that meaning is use (in fact, I take this thesis to be a contribution to essentially fight such a simplistic and in my view wrong-headed theory of meaning). It simply means that we should refrain from too much arm-chair speculation and assumptions of meaning merely backed up by the tradition but not by actual empirical evidence. It means that the meaning of our terms may not be conflated with that which we use to decide whether a term applies or not.

This confusion between empirical issues of how we actually apply our concepts and more conceptual normative issues turn out to be not only difficult to elaborate, but also going against deep-rooted assumptions in both philosophy and psychology, as well as linguistics, that the meaning of a term is a concept that determines the right application of this term. Assuming that competent speakers are able to apply at least some words correctly, they must therefore possess the respective meaning or concept. I have argued that this approach is misguided, which however opens up potential to study and theorize about how we in fact apply words. Most

importantly, it opens an attractive way of bringing results from philosophy, psychology and linguistics together into a single coherent framework, which has been one important aim of my work.

Bibliography

Adams, F. and Campbell, K. (1999). Perceptual symbol systems. *Behavioral Brain Sciences*, 22: 610-611.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral Brain Sciences*, 22:577-660.

Barsalou, L. W. (2012). The human conceptual system. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 239–258). Cambridge: Cambridge University Press.

Bergen, B. K. (2012). *Louder than words: The new science of how the mind makes meaning*. NewYork: Basic Books.

Bermúdez, J. L. (2004). *Philosophy of psychology: A contemporary introduction*. New York: Routledge.

Borg, E. (2012). P*ursuing meaning*. Oxford: Oxford University Press.

Borghi, A. M. and Cimatti, F. (2009). Words as tools and the problem of abstract words meanings. In: *Proceedings of the 31st annual conference of the cognitive science society*, volume 31, pages 2304-2309.

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe?. *Philosophical studies*, *170*(3), 465-500.

Boyd, R., (1999). Homeostasis, species, and higher taxa. In R. Wilson (ed.), S*pecies: New Interdisciplinary Essays*, 141-185. Cambridge: MIT Press.

Bruno, N. (2001) When does action resist visual illusions? *Trends in Cognitive Sciences*, 5, 385–8.

Burge, T. (1979). Individualism and the Mental. *Midwest studies in philosophy*, 4(1): 73–121.

Burr, V. (2015). *Social constructionism*. New York: Routledge.

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

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge University Press.

Carston, R. (2008). *Thoughts and utterances: The pragmatics of explicit communication*. John Wiley & Sons.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press.

Clark, A. and Prinz, J. J. (2004). Putting Concepts to Work: Some Thoughts for the Twenty-First Century. *Mind and Language*, 19(1):57-69.

Collins, J. (2009). II—John Collins: Methodology, Not Metaphysics: Against Semantic Externalism. In *Aristotelian Society Supplementary Volume* (Vol. 83, No. 1, pp. 53-69). Oxford, UK: Oxford University Press.

Cruse, A., (1999). *Meaning in Language: An introduction to Semantics and Pragmatics*. New York: Oxford University Press.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural computation*, 1(1), 123-132.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology*, 122(3), 371-396.

Del Pinal, G. (2015). Dual content semantics, privative adjectives and dynamic compositionality. *Semantics & Pragmatics*: 8 (7): 1–53.

Del Pinal, G. (2016). Prototypes as compositional components of concepts. *Synthese*, 193(9), 2899–2927.

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

Diaz-León, E. (2015). What is social construction?. *European Journal of Philosophy*, 23(4): 1137–1152.

Dove, G. (2009). Beyond perceptual symbols: a call for representational pluralism. *Cognition*, 110:412-431.

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, 1.

Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic bulletin & review*, 23(4), 1109-1121.

Dummett, M. (1993). *The seas of language* (pp. 160-162). Oxford: Clarendon Press.

Edwards, K. (2010). Unity amidst heterogeneity in theories of concepts. *Behavioral and Brain Sciences*, 33(2–3), 210–211.

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. *Journal of verbal learning and verbal behavior*, *21*(3), 296-306.

Evans, G. (1973). The Causal Theory of Names. *Proceedings of the Aristotelian Society*, Supplementary Volume 47: 187–208.

Fias, W., & Fischer, M. (2005). Spatial representation of numbers. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 43-54). New York, NY: Psychology Press.

Fodor, J. A., (1989). *Why there still has to be a language of thought*. In: Computers, Brains and Minds (pp. 23–46). Springer Netherlands.

Fodor, J. A., (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Fodor, J. (2004). Having concepts: A brief refutation of the twentieth century. *Mind and Language*, 19(1), 29–47.

Fodor, J. A., & Lepore, E. (1992). *Holism: A shopper's guide*. Chichester, U.K.: Wiley.

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition*, *58*(2), 253-270.

Frege, G., (1948). Sense and reference. *The philosophical review*, 57(3), 209–230.

Gallagher, S. (2008). Inference or Interaction: Social Cognition without Precursors. *Philosophical Explorations*. 11(3):163-74.

Gallese V. & Lakoff G (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3/4): 455-479.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12), 493-501.

Gärdenfors, P. (2004). *Conceptual spaces*: The geometry of thought. MIT press.

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. New York: Oxford University Press.

Gernsbacher, M. A. (1995). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Glasgow, J. (2009). *A Theory of Race*. New York: Routledge.

Goodale, M. A. and Humphrey, G. K. (1998) The objects of action and perception. *Cognition*, 67, 181–207.

Gotham, M. G. H. (2014). *Copredication, quantification and individuation* (Doctoral dissertation, UCL (University College London)).

Guala, F. (2016). *Understanding institutions: The science and philosophy of living together*. New Jersey: Princeton University Press.

Hacking, I. (1999). *The social construction of what*?. Cambridge, Mass.: Harvard University Press.

Hampton, J. A. (2000). Concepts and prototypes. *Mind & Language*, 15(2–3), 299–307.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of Learning and Motivation*, 46, 79–113.

Hampton, J. A., & Jönsson, M. L. (2012). Typicality and Composition a Lity: the Logic of Combining Vague Concepts. In: Werning, W. Hinzen, and E. Machery (eds.). *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Haslanger, S. (2005). What are we talking about? The semantics and politics of social kinds. *Hypatia*, 20(4): 10–26.

Horn, L. (2017). Pragmatics and the Lexicon. In: Huang, Y. (edt.). *The Oxford handbook of pragmatics*. Oxford: Oxford University Press. 511–532.

Kenny, a. (2010). Concepts, brains, and behaviour. *Grazer philosophische studien*, *81*(1).

Lewis, D. (1970). General semantics. *Synthese*, 22, 18–67.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford University Press.

Jones, M. (2015). Number concepts for the concept empiricist. *Philosophical Psychology*, 29 (3):334-348.

Keil, F., (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Khalidi, M. A. (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge: Cambridge University Press.

Khalidi, M. A. (2015). Three kinds of social kinds. *Philosophy and Phenomenological Research*, 90(1): 96–112.

Khalidi, M. A. (2016). *Natural Kinds. In: The Oxford Handbook of Philosophy of Science*. Oxford University Press.

Kiefer, M. and Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, 48(7):805-825.

Kite, M. E., & Whitley Jr., B. E. (2016). *Psychology of prejudice and discrimination*. New York: Routledge.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127(2), 242-257.

Kripke, S. A. (1972). *Naming and necessity*. Cambridge, Mass.: Harvard University Press.

Lakoff, G. (1987). *Women, Fire and Dangerous Things. What Categories Reveal about the Mind*. Chicago: Chicago University Press.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. Chicago: Chicago University Press.

Lalumera, E., (2010). Concepts are a functional kind. *Behavioral and Brain Sciences*, 33(2–3), 217-218.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, 86(1), 25-55.

Laurence, S., & Margolis, E., (1999). Concepts and cognitive science. In: Laurence, S., & Margolis, E. (eds.). *Concepts: core readings*, 3–81. Cambridge, MA: MIT Press.

Lewis, D. (1970). General Semantics. *Synthese* 22: 18-67.

Liebesman, D., & Magidor, O. (2017). Copredication and property inheritance. *Philosophical Issues*, *27*(1), 131-166.

Lindblom, J. (2015). *Embodied social cognition* (Vol. 26). Heidelberg: Springer.

Lindemann, O., & Fischer, M. H. (2015). Embodied number processing. *Journal of Cognitive Psychology*, 27(4), 381-387.

Liptow, J. (2012). Thinking and Judging. *Grazer Philosophische Studien*, 86(1), 223-234.

Löhr, G. (2017). Abstract concepts, compositionality, and the contextualism-invariantism debate. *Philosophical Psychology*, 30(6), 689–710.

Löhr, G. (2018). Concepts and categorization: do philosophers and psychologists theorize about different things? *Synthese*. Online first.

Löhr, G. (2019). Embodied cognition and abstract concepts: Do concept empiricists leave anything out?. *Philosophical Psychology*, 32(2), 161-185.

Machery, E. (2005). Concepts are not a natural kind. *Philosophy of Science*, 72(3), 444–467.

Machery, E. (2006). Two Dogmas of Concept Empiricism. *Philosophy Compass*, 1(4):398-412.

Machery, E. (2007). Concept empiricism: A methodological critique. *Cognition*, 104(1), 19-46.

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

Machery, E., (2010a). Précis of doing without concepts. *Behavioral and Brain Sciences*, 33(2–3), 195–206.

Machery, E., (2010b). Replies to my critics. *Philosophical Studies*, 149(3), 429–436.

Machery, E., (2011). Variation in intuitions about reference and ontological disagreement. In: S. D. Hales (Ed.). *A Companion to Relativism* (pp. 118–136). Malden, MA: Wiley-Blackwell.

Machery, E. (2014). Social Ontology and the Objection from Reification. In: M. Gallotti and J. Michael. *Perspectives on Social Ontology and Social* Cognition. Dordrecht: Springer: 87–102.

Machery, E. (2015). By Default: Concepts Are Accessed in a Context-Independent Manner. In: Margolis, L. & Laurence, S. (eds). *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press.

Machery, E. (2016). The amodal brain and the offloading hypothesis. *Psychonomic bulletin & review*, 23(4), 1090-1095.

Machery, E. (2017). *Philosophy within its proper bounds*. Oxford: Oxford University Press.

Machery, E. & Lederer, L., (2012). Simple Heuristics for Concept Combination. In: Werning, W. Hinzen, and E. Machery (Eds.). *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press, (pp. 81–106).

Magliano, J. P., Baggett, W. B., Johnson, B. K., & Graesser, A. C. (1993). The time course of generating causal antecedent and causal consequence inferences. *Discourse Processes*, *16*(1-2), 35-53.

Mahon, B. Z. and 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):59-70.

Mallon, R. (2003). Social construction, social roles, and stability. In F. Schmitt (ed.), Socializing Metaphysics: *The Nature of Social Reality*. Rowman & Littlefield: 65-91.

Mallon, R. (2004). Passing, traveling and reality: Social constructionism and the metaphysics of race. *Noûs*, 38(4), 644-673.

Mallon, R. (2014). "Naturalistic Approaches to Social Construction", The Stanford Encyclopedia of Philosophy (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2014/entries/social-construction-naturalistic/>.

Mallon, R. (2016). *The construction of human kinds*. Oxford University Press.

Mallon, R. (2017). Social Construction and Achieving Reference. *Noûs*, 51(1), 113-131.

Malt, B. C. (1994). Water is not H2O. *Cognitive psychology*, 27(1), 41-70.

Margolis, E., (1998). How to acquire a concept. *Mind & Language*, 13(3), 347–369.

Margolis, E., & Laurence, S. (2010). Concepts and theoretical unification. *Behavioral and Brain Sciences*, 33(2–3), 219–220.

Margolis, E., & Laurence, S., (2011). Learning matters: The role of learning in concept acquisition. *Mind & Language*, 26(5), 507–539.

Margolis, E., & Laurence, S. (2014). Concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/spr2014/entries/concepts/>.

Margolis, E., & Laurence, S. (2019). Concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, <https://plato.stanford.edu/archives/sum2019/entries/concepts/>.

Mazzone, M., & Lalumera, E., (2010). Concepts: Stored or created? *Minds and Machines*, 20, 47–68.

Medin, D. and Ortony, A. (1989). Psychological Essentialism. In S. Vosniadou and A. Ortony (eds), *Similarity and Analogical Reasoning*. Cambridge University Press: 179–95.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming off age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7), 788-804.

Michael, J. (2017). Putting unicepts to work: A teleosemantic perspective on the infant mindreading puzzle. *Synthese*, 194(11), 4365–4388.

Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More Mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21(01), 55–65.

Millikan, R. G. (2017). *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.

Mitchell, R. and Clement, A. (1999). Perceptual symbol systems. *Behavioral Brain Sciences*, 22:628-629.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, 22(2): 103–122.

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

Nimtz, C., & Langkau, J., (2010). Concepts in Philosophy-A Rough Geography. *Grazer Philosophische Studien*, 81(1), 1.

Ohlsson, S. (1999). Perceptual symbol systems. *Behavioral Brain Sciences*, 22:630-631.

Pagin, P. (2012). Communication and the complexity of semantics. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality*. Oxford: Oxford University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

Peacocke, C. (1989). What Are Concepts? 1. *Midwest Studies in Philosophy*, *14*(1), 1-28.

Peacocke, C., (1992). *A study of concepts*. The MIT Press.

Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010). Amodal semantic representations depend on both left and right anterior temporal lobes: New rTMS evidence. *Neuropsychologia*, 48, 1336-1342.

Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Prinz, J. J. (2005). The return of concept empiricism. In: Cohen, H., & Lefebvre, C. (Eds.). *Handbook of categorization in cognitive science*, 679-699. Elsevier.

Prinz, J. (2012). Regaining composure: A defense of prototype compositionality. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality*. Oxford: Oxford University Press.

Prinz, J., & Clark, A. (2004). Putting concepts to work: Some thoughts for the twentyfirst century. *Mind & Language*, 19(1), 57-69.

Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, 17(9), 458-470.

Pulvermüller, F., Hauk, O., Nikulin, V. & Ilmoniemi, R.J. (2005). Functional interaction of language and action: a TMS study. *European Journal of Neuroscience*, 21 (3), 793-797.

Putnam, H. (1975). The meaning of'meaning'. In Pessing, A, G. S., editor*, The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's the meaning of meaning*. London: Routledge: 3–52.

Putnam, H., (1973). Meaning and reference. *The journal of philosophy*, 70(19), 699–711.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(01):111-132.

Quine, W. (1951). Two Dogmas of Empiricism, *Philosophical Review*, 60: 20-43.

Recanati, F. (2010). *Truth-conditional pragmatics*. Oxford: Oxford University Press.

Recanati, F. (2012). *Mental files*. Oxford: Oxford University Press.

Rey, G. (2010). Concepts versus conceptions (again). *Behavioral and Brain Sciences*, 33(2–3), 221–222.

Rey, G., (1983). Concepts and stereotypes. *Cognition*, 15(1), 237–262.

Rey, G., (1985). Concepts and conceptions: A reply to Smith, Medin and Rips. *Cognition*, 19(3), 297–303.

Rey, G., (2009). Review of E. Machery, Doing without Concepts. *Notre Dame Philosophical Reviews.* (Online journal. Epub: 2009.07.15.) Available at http:// ndpr.nd.edu/review.cfm?id1⁄416608

Rice, C. (2013). Concept empiricism, content, and compositionality. *Philosophical studies,* 162(3), 567-583.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization. Hillsdale: Erlbaum. Reprinted in: Margolis, E. and Laurence, S. (Eds.) (1999). *Concepts: Core readings*. Cambridge: MIT Press.

Rosch, E., (1983). Prototype classification and logical classification: The two systems. New trends in conceptual representation: *Challenges to Piaget's theory*, 73–86.

Rosch, E. H. (2011). Slow lettuce: categories, concepts, fuzzy sets, and logical deduction. In: Belohlavek, R. and Klir, G. (eds.). *Concepts and fuzzy logic*. MIT Press, 89–120.

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Searle, J. R. (1958). Proper names. *Mind*, 67(266): 166–173.

Searle, J. R. (1995). *The construction of social reality*. New York: Simon and Schuster.

Siebel, M. (1999). Perceptual symbol systems. *Behavioral Brain Sciences*, 22:632-633.

Smith, E. E., Medin, D. L., & Rips, L. J. (1984). A psychological approach to concepts: Comments on Rey's "concepts and stereotypes". *Cognition*, 17(3), 265–274.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions on word meaning: object terms and substance terms. *Cognition*, 38: 179–211.

Spelke, E. S., Bernier, E. P., & Skerry, A. E. (2013). Core social cognition. In: Banaji, M. R., & Gelman, S. A. (Eds.). *Navigating the social world. What infants, children, and other species can teach us.* Oxford: Oxford University Press: 11–16.

Stalnaker, R. (1997). Reference and Necessity. In B. Hale and C. Wright (Eds.), *A Companion to the Philosophy of Language* (pp. 534-54). Oxford: Blackwell.

Szabó, Z. (2012). The case for compositionality. In: Werning, Markus, Wofram Hinzen, and Edouard Machery (eds.) *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press. 64–80.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, 44(4): 929–938.

Tomasello, M. (2009). *The cultural origins of human cognition*. Cambridge, Mass.: Harvard University Press.

Travis, C. (2008). *Occasion-sensitivity: Selected essays*. Oxford University Press.

Vakkuri, J. (2018). Semantic externalism without thought experiments. *Analysis*, 78(1), 81–89.

Vicente, A. (2018) Polysemy and word meaning: an account of lexical meaning for different kinds of content words, *Philosophical Studies,* 175 (4), 947-968

Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., and Cappa, S. F. (2014). The neural representation of abstract words: the role of emotion. *Cerebral Cortex*, 24(7):1767-1777.

Weiskopf, D. A. (2009). The plurality of concepts. *Synthese*, *169*(1), 145.

Werning, M. (2005). Right and wrong reasons for compositionality. In: Werning, M., Machery, E., & Schurz, G. (eds.). *The compositionality of meaning and content,* 1, 285–309.

Wilson-Mendenhall, C. D., Simmons, W. K., Martin, A., & Barsalou, L. W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of cognitive neuroscience*, 25(6), 920-935.

Wittgenstein, L. (1953/2009). *Philosophical investigations*. John Wiley & Sons.

Wynne, C. D., (2001). *Animal cognition: The mental lives of animals*. Macmillan.

Yli-Vakkuri, J. (2018). Semantic externalism without thought experiments. *Analysis*, 78(1), 81–89.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 386-397.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic bulletin & review*, *23*(4), 1028-1034.