UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE FILOSOFIA
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

RAQUEL ALBIERI KREMPEL

An essay on the language of thought
Um ensaio sobre a linguagem do pensamento

Versão corrigida

São Paulo
2018

RAQUEL ALBIERI KREMPEL

An essay on the language of thought
Um ensaio sobre a linguagem do pensamento

Versão corrigida

Tese de Doutorado apresentada ao Programa de Pós-graduação em Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas, da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Doutora em Filosofia.

Orientador: Prof. Dr. João Vergílio Gallerani Cuter

São Paulo
2018

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

K92a

Krempel, Raquel
An essay on the language of thought / Raquel Krempel ; orientador João Vergílio Cuter. - São Paulo, 2018.
255 f.

Tese (Doutorado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Filosofia. Área de concentração: Filosofia.

1. Pensamento. 2. Linguagem. 3. Fodor. 4. Conceitos. I. Cuter, João Vergílio, orient. II. Título.

KREMPEL, R. A. **An essay on the language of thought/Um ensaio sobre a linguagem do pensamento**. Tese apresentada à Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo para a obtenção do título de Doutora em Filosofia.

Aprovada em:

Banca Examinadora

Prof. Dr. _____ Instituição:_____

Julgamento:_____ Assinatura:_____

Prof. Dr. _____ Instituição:_____

Julgamento:_____ Assinatura:_____

Prof. Dr. _____ Instituição:_____

Julgamento:_____ Assinatura:_____

Prof. Dr. _____ Instituição:_____

Julgamento:_____ Assinatura:_____

Prof. Dr. _____ Instituição:_____

Julgamento:_____ Assinatura:_____

To the memory of Jerry Fodor.

ACKNOWLEDGEMENTS

I thank my supervisor João Vergílio Gallerani Cuter, for so many years of supervision and encouragement. No other Wittgensteinian would have been so supportive of a work on the language of thought. I also thank professors Osvaldo Pessoa and Marcos Lopes for their helpful comments during the preliminary examination. I also thank them all for teaching me so much in all the years (more than I like to admit) I have been a student at the University of São Paulo.

I thank professor Jerry Fodor, for kindly accepting me as a visiting student at Rutgers, at the beginning of this project. I thank him, above all, for showing me that philosophy can be worth doing. Professor Peter Carruthers was also kind enough to accept me as a visiting student at the University of Maryland. I thank him for that, and for guiding me into making parts of chapter 4 not completely embarrassing. I also thank professor Eve Danziger, for kindly allowing me to sit in her course on Language and Thought. I learned a lot from her.

Parts of this dissertation were presented at a group with my supervisor João Vergílio and his students. I thank them for their suggestions, which helped to improve this dissertation. I also had the opportunity to discuss parts of this dissertation in the meetings of GEMF (Grupo de escrita de mulheres na filosofia). Thanks to Beatriz Sorrentino, Eduarda Calado, Larissa Gondim and Nara Figueiredo for their friendship and the fruitful philosophical discussions.

My friends Nathalie Bressiani, Nara Figueiredo, Renata Itagyba, Patrícia Ribeiro de Lima and José Wilson Silva helped in this process more than they probably know. Special thanks to Patrícia, for helping me *format* the dissertation.

I also thank my Brazilian, American and feline families, for their support. In particular Miguelinho, Lucy and my mother (in that order).

This dissertation wouldn't have existed without Evan's love, support, encouragement and willingness to debate the language of thought. I can never thank him enough. But if there are any English mistakes left, they are completely his fault. I will take the blame for the mistakes in the thoughts being expressed.

Many thanks also to the secretaries of the Philosophy Department, for all the help.

This research was supported by grants # 2014/15037-0 and # 2017/02074-2, São Paulo Research Foundation (FAPESP) and by CAPES.

The conceptual term is an intention or impression of the soul which signifies or consignifies something naturally and is capable of being a part of mental proposition and of suppositing in such a proposition for the thing it signifies. Thus, these conceptual terms and the propositions composed of them are the mental words which, according to St. Augustine in chapter 15 of *De Trinitate*, belong to no language. They reside in the intellect alone and are incapable of being uttered aloud, although the spoken words which are subordinated to them as signs are uttered aloud.

(WILLIAM OF OCKHAM)

RESUMO

KREMPEL, R. A. **Um ensaio sobre a linguagem do pensamento**. Tese (Doutorado), Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2018.

O objetivo geral desta tese é esclarecer e discutir diversos tópicos relacionados, de um modo ou de outro, à hipótese da linguagem do pensamento, formulada de maneira mais elaborada por Jerry Fodor. A hipótese da linguagem do pensamento é uma hipótese sobre a natureza das representações mentais. Ela diz que representações mentais têm uma estrutura linguística. Isso é o mesmo que dizer que, tal como sentenças em uma língua natural, representações mentais têm constituintes primitivos (com propriedades sintáticas e semânticas), que se combinam para formar símbolos sintática e semanticamente complexos. A hipótese da linguagem do pensamento está profundamente relacionada à teoria representacional e à teoria computacional da mente. Discutirei essas teorias e as compararei com algumas visões filosóficas opostas da mente. Em seguida, discutirei os argumentos da produtividade e da sistematicidade, em favor da linguagem do pensamento. Finalmente, veremos diferentes modos de conceber a relação entre a linguagem do pensamento e as línguas naturais.

Palavras-chave: Pensamento. Linguagem. Fodor. Conceitos.

ABSTRACT

KREMPEL, R. A. **An essay on the language of thought**. Tese (Doutorado), Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2018.

The general goal of this dissertation is to clarify and discuss several topics that are, in some way or other, related to the language of thought hypothesis, put forward most forcefully by Jerry Fodor. The language of thought hypothesis is a hypothesis about the nature of mental representations. It says that mental representations have a linguistic structure. This is the same as saying that, just like sentences in a natural language, mental representations have primitive constituents (with syntactic and semantic properties), which combine to form syntactically and semantically complex symbols. The language of thought hypothesis is closely related to the representational and computational theories of mind. I discuss them and compare them to some opposing philosophical views of the mind. I then discuss the productivity and the systematicity arguments for the language of thought. Finally, we will see different ways to conceive the relation between the language of thought and the natural languages.

Keywords: Thought. Language. Fodor. Concepts.

ABBREVIATIONS

LOT	Language of thought
RTM	Representational theory of mind
CTM	Computational theory of mind

CONTENTS

INTRODUCTION

The general goal of this dissertation is to clarify and discuss several topics that are, in some way or other, related to the language of thought hypothesis, put forward most forcefully by Jerry Fodor (1975). The language of thought hypothesis is a hypothesis about the nature of mental representations. It says that mental representations have a linguistic structure. This is the same as saying that, just like sentences in a natural language, mental representations have primitive constituents (with syntactic and semantic properties), which combine to form syntactically and semantically complex symbols. It is usually accepted by its proponents that the language we think in is not to be identified with any natural language, such as English or Portuguese.

There are some general views of the mind that usually come together with the language of thought hypothesis. In chapter 1, my main goal is to introduce them. To the extent that the language of thought hypothesis says that linguistically structured mental representations are the vehicle of thought, it presupposes that there are mental representations. This is why I will first introduce the representational theory of mind. Also, the assumption that we think in a language of thought is usually presupposed by a computational theory of mind, according to which mental processes are computational processes. So we will also look at some of its main features. In the end of chapter 1, I will discuss some of the limitations of the representational and computational theories of mind, as conceived by Jerry Fodor. It will be important to keep in mind that the language of thought is not supposed to provide a complete explanation of mental states. It is, above all, a view about the mental representations that appear in propositional attitudes.

Jerry Fodor will be my main interlocutor throughout this dissertation. We owe to him the best formulation of the language of thought hypothesis, and of several issues related to it. I will try both to clarify some of his views about the language of thought and the representational and computational theories of mind of which it is part, and to raise some problems for them. But in order to better understand the language of thought and the representational and computational theories of mind, it might help to consider some opposing views. In chapter 2, I compare the representational theory of mind with some reductionist philosophical views about the mind: behaviorism and the identity theory. The representational theory of mind, as we will see, has the

advantage of admitting states with the same properties that common sense psychology attributes to intentional mental states: semantic evaluability and causal efficacy. I will also discuss the multiple realizability argument, which is usually introduced against the identity theory, and as a way of arguing for the autonomy of psychological states and explanations. Finally, I will discuss some of Searle's criticisms of computational accounts of the mind, by imagining an exchange between Searle and Fodor.

In chapter 3, we will see the productivity and the systematicity arguments for the existence of a language of thought. I will also briefly present some of Fodor, Pylyshyn and McLaughlin's criticisms against connectionism, conceived as a cognitive model of the mind. Finally, I will discuss one argument Fodor formulated against the compositionality of natural languages, with which he hoped to show that the content of thought is prior to the content of language. I will argue that Fodor's argument is circular, and that it conflicts with his previous arguments for the language of thought.

Though the productivity and systematicity arguments support the view that the vehicle of thought is a language, they are neutral as to whether the language we think in is the same or different from any natural language. In chapter 4, I will introduce some arguments (put forward by Fodor, Pinker and Pylyshyn) for the idea that the language of thought is different from any natural language. I will also discuss some of Carruthers' and Fodor's ideas about the relation between thought and language. In order to challenge some of Fodor's views about concepts and their relation with words, I will take into consideration some empirical findings about color words and their influence on other cognitive domains. I will indicate that these findings may support an intermediate position – in between Carruthers' and Fodor's – about the role of natural language on thought. It is commonly assumed that, in holding that there is a language of thought, one is thereby committed to the view that the language of thought is innate and universally shared. So if one were to show that the natural language one speaks influences thought, that would be an argument against the view that we think in a language of thought different from any natural language. At the end of this final chapter, I will indicate that this is a misconception. It is not essential to the language of thought that it be innate and universal. So even if it is true that one's natural language has some influence on thought, it can still be the case that the vehicle of thought is a non-natural language.

Though I will be discussing several topics related to the language of thought hypothesis, I am not offering an exhaustive treatment of them. On the contrary, several important issues will be only briefly, if at all, touched in this dissertation. These include the semantics and individuation of concepts (the primitive symbols of the language of thought), the question of whether we also employ representations that are not linguistically structured (e.g. mental images and maps), whether animals also have a language of thought, whether (assuming we accept the modularity thesis) we have one language of thought for each module in the mind, and so on. I hope to deal with them in future research.

CHAPTER 1

THE REPRESENTATIONAL AND COMPUTATIONAL THEORIES OF MIND

> "The form of a philosophical theory, often enough, is: *Let's try looking over here*."
>
> (Fodor, 1981c)

Jerry Fodor famously argues that thought occurs in a language, which is different from any natural language, and which he calls the language of thought (LOT). The main goal of this dissertation is to clarify and discuss some issues related to the language of thought hypothesis. First, though, it is important to clarify some theories of the mind that are closely related to it. In the first part of this chapter, we will see the representational (RTM) and computational theories of mind (CTM). We will see that these theories should not be taken as providing a complete picture of the mind. Rather, a reasonable way to conceive them, the one that Fodor originally proposes, is as vindicating at least part of common sense psychology, in a naturalistic way. With RTM, Fodor wishes to provide a naturalistic conception of propositional attitudes, that is, mental states like beliefs and desires. According to him, being in one of these states is being in a functional relation with a mental representation. As for CTM, it provides a naturalistic way of conceiving mental processes, such as reasoning. According to Fodor, at least some mental processes can be characterized as transformations of symbols which are driven solely by their syntax, and not by their semantic content.

As we will see in the first part of the chapter, Fodor characterizes propositional attitudes in part by saying that they are *functional* states. This raises the question of whether Fodor can be considered a proponent of functionalism about the mind. Though it is quite common to do so, in the second part of this chapter we will see that Fodor actually criticizes two forms of functionalism: machine state functionalism and semantic functionalism (or functional-role semantics). I will try to clarify the extent to which Fodor can be characterized as a functionalist. His view is that attitudes, such as believing and desiring, can be individuated by their functions. Particular propositional attitudes, on the other hand, cannot be individuated by their functions or causal roles. Fodor's view about the mind can be described as a moderate functionalism, to which

is added the assumption that mental states involve representation (RTM) and that mental processes are computational processes (CTM).

In the third part of the chapter, I will present some limitations of RTM and CTM. First, I will show that RTM doesn't vindicate common sense psychology completely, because it doesn't explain mental states with a qualitative character. Second, I will indicate that the scope of the computational theory of mind, as a theory of mental processes, is also limited. Though it is important to keep these limitations in mind, they don't speak against RTM as a theory of propositional attitudes, or against CTM as a theory of demonstrative mental processes.

1. LOT, RTM, CTM

The language of thought hypothesis, as formulated by Fodor, is a hypothesis about the nature of mental representations. It is part of his attempt to formulate a theory that explains some aspects of the mind. What it says is that mental representations have a linguistic structure. This is the same as saying that, just like sentences in a natural language, mental representations have primitive constituents (with syntactic and semantic properties), which combine to form syntactically and semantically complex symbols. This hypothesis clearly presupposes the adoption of a representational theory of mind, given that it is committed to the existence of mental representations. The representational theory of mind offered by Fodor is a theory about the nature of propositional attitudes, that is, about intentional mental states such as beliefs and desires.[1] Once we assume that there are mental representations, and that these representations have a structure, we can give an extra step and adopt the computational theory of mind, which says that mental processes, such as reasoning, are computational processes. In what follows I will try to clarify these two theories (RTM and CTM) and what motivates them.

[1] States like beliefs, desires, intentions, guesses, fears, regrets, etc. are called propositional attitudes because it is generally assumed that they are different attitudes one can take about a proposition. It is also common to call them intentional states, in the sense that what is essential to them is that they are about something. We could naturally raise questions about the nature of this "something" beliefs and desires are about. Is the object of intentional states always a proposition, or can some intentional states at least sometimes be attitudes toward objects? According to Searle, intentional states can be either about states of affairs or about objects. Fodor, as far as I know, does not discuss this issue. For my part, I agree with Searle that it is possible to have mental states that are directed towards objects or properties, and not propositions (for instance, I can like chocolate, but hate vanilla). But this discussion will not be addressed here. I will simply use "propositional attitudes" interchangeably with "intentional states". I also won't be dealing with the difficult issue of determining what a proposition is. In any case, what will be discussed here will not depend on these subtleties.

In *Psychosemantics*, Fodor makes it clear that he wants to develop a naturalistic theory of the mind which vindicates common sense psychology. That is to say, he wishes to provide a theory which is compatible with the natural sciences, and which accepts entities with the essential characteristics that common sense attributes to mental states such as beliefs and desires. He believes that his representational and computational theories of mind allow one to understand, in a naturalistic way, how mental states can have two characteristics that we commonly attribute to them: causal efficacy and semantic evaluability.[2] Let us first look at the idea of causal efficacy. Common sense seems to suppose the existence of three types of mental causation: mental states cause behaviors, external events cause mental states, and mental states cause other mental states. As for the first kind of mental causation, we often explain the behavior of other people by appealing to their mental states, such as beliefs and desires, that we assume cause them. We say, for example, that John went to the bakery because he *believed* that there they sell bread, and because he *wanted* to eat bread. In this case, we are saying that John's beliefs and desires caused his behavior.[3] The second type of mental causation, that is, the assumption that external events can cause mental states, also appears frequently in our everyday psychological explanations. I usually suppose, for example, that the noise coming from car horns gives me a headache and the desire to move out of São Paulo. Finally, we normally assume that mental states cause other mental states, which is what we call mental processes. The thought that it will rain, together with my desire not to get wet, may cause the thought that it is better to stay at home, and eventually my decision to stay home. In short, our daily psychological explanations of behavior assume that mental states are causally efficacious.[4]

Moreover, according to Fodor, common sense also considers that the same mental states that have causal efficacy also have semantic content, which makes them semantically evaluable. That is, mental states such as beliefs, desires, hunches, etc., have a relation to the non-psychological world that makes them true or false (in the case of beliefs), fulfilled or frustrated (in

---

[2] Another characteristic of common-sense psychology that a naturalistic theory must vindicate, according to Fodor, is the implicit generalizations we use, for example, to predict and explain other people's behavior. A recurrent example in his books is: "if *x* wants that *P*, and *x* believes that not-*P* unless *Q*, and *x* believes that *x* can bring it about that *Q*, then (ceteris paribus) *x* tries to bring it about that *Q*." (FODOR, 1987, p. 02).

[3] Fodor does not find the distinction between causes and reasons relevant here. For him, the way we explain behavior often enough gives us the causes of behavior (cf. FODOR, 1987, p. 20).

[4] We shall see later that although Fodor notes that there are all these types of mental causation, the central type that his computational theory of mind intends to explain is the third, that is, mental processes. This is because Fodor is interested in providing a purely cognitive theory of the mind, and the other types of mental causation presumably would require explanations that appeal not only to cognitive aspects but also to physiological aspects.

the case of desires), or right or wrong (in the case of hunches) (cf. FODOR, 1987, p. 10). A belief, for instance, has a certain content, or is about something, or represents the world as being in a certain way (or expresses a given proposition), and this makes it semanticaly evaluable, that is, capable of being true or false.[5]

Fodor wants a theory that vindicates common sense psychology mainly because, in his usual alarmist style, he believes that if it is false, "that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species; if we're that wrong about the mind, then that's the wrongest we've ever been about anything." (1987, p. xii). As Fodor famously says, psychological explanations are so pervasive in our lives that, were we to find that it is not literally true that mental states have causal powers, this would be the end of the world.[6] Our everyday explanations of behavior usually work, and Fodor argues that this is a sign that they are for the most part true. What we need is a theory that shows us how common sense psychology can be true.

One of the difficulties for a theory of mind that is intended to vindicate common-sense psychology is to explain how states with semantic content can have causal efficacy.[7] How can beliefs and desires, which seem to be essentially intentional states (in the sense of being *about* things), cause and be caused by other mental states, external events, and behaviors? How can, say, my belief that it is raining cause me to get an umbrella? For Fodor, the only psychological theories that vindicate common-sense psychology, which give a naturalistic account of mental states taken to be at once causally efficacious and semantically evaluable, without eliminating or reducing them to entities of more basic scientific theories, are the representational and the computational theories of the mind.[8] According to him,

---

[5] Or, as Searle would say, intentional states such as beliefs and desires have conditions of satisfaction. For Searle, however, not all states with semantic content – which are characterized by being directed at or by being about objects or states of affairs in the world – are states that have conditions of satisfaction. Me being glad because my friend was awarded a prize, or upset because I insulted someone, as Searle notes, do not seem to be states that are satisfied or not by some relation to the world, in the way that a belief is satisfied if it is true, or not satisfied if it is false (cf. SEARLE, 1983, p. 08). What is important to note is that even if not all propositional attitudes can be said to be semantically evaluable, Fodor would certainly say that they all have semantic content, in the sense of being about something.

[6] In Fodor's words, "if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying ..., if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world." (1989, p. 156).

[7] As Crane puts it, "if we are going to explain thought, then we have to explain how there can be states which can at the same time be representations of the world and causes of behavior." (2003, p. 83).

[8] It is not always clear whether Fodor considers RTM and CTM psychological theories, or philosophical theories about the mind that are compatible with the theories of cognitive psychology. In *The Language of Thought*, Fodor

> What's important about RTM (…) [is] that RTM shows how intentional states could have causal powers; precisely the aspect of commonsense intentional realism that seemed most perplexing from a metaphysical point of view. (FODOR, 1987, p. 26).
>
> the full-blown Representational Theory of Mind (…) purports to explain how there *could be* states that have the semantical and causal properties that propositional attitudes are commonsensically supposed to have. In effect, RTM proposes an account of what the propositional attitudes *are*. (FODOR, 1990, p. 05).

Let us briefly look first at what RTM says about intentional mental states, and then at what CTM says about mental processes. According to Fodor, propositional attitudes such as beliefs and desires are functional relations that organisms have with mental representations. If I have a belief or a desire, what happens is that I have a mental representation to which I am related in a certain way, which varies depending on whether my mental state is a belief, a desire, etc. Fodor adopts a metaphor proposed by Steven Schiffer, and says that the difference between believing *P* and desiring *P* is the difference between having a token of the mental representation *P* in a "belief box," or a "desire box." This is just "a shorthand for representing the attitudes as *functional* states" (FODOR, 1987, p. 17). According to Fodor, "the difference between having the token [of a mental representation] in one box or the other corresponds to the difference between the causal roles of beliefs and desires." (idem). In addition, "to be – metaphorically speaking – in the state of having such and such a rock in your intention box is just to be – literally speaking – in a state that is the normal cause of certain sorts of effects and/or the normal effect of certain sorts of causes." (1987, p. 138).

Thus having the belief that it will rain, for example, is to be in a functional relation with a mental representation or symbol that means that it will rain. To wish or hope for rain is to be in a different functional relation, with different causal roles, with the same mental symbol. Fodor adopts, in a way, a functionalist position on the characterization of propositional attitudes. Beliefs

---

describes his own work as speculative psychology, because "it is an attempt to say how the mind works insofar as answers to that question emerge from recent empirical studies of language and cognition." (1975, p. viii) In *Psychosemantics*, he calls RTM "a (more or less) empirical theory." (1987, p. 16). More likely, then, RTM and CTM are a bit of both. Many philosophers have given up the attempt to provide a demarcation between philosophy and the sciences, and Fodor is clearly one of them. As I see the issue, philosophers of mind usually raise broader, and more fundamental questions about the subject matter of psychology, without typically going to the trouble of designing experiments, but sometimes being constrained by empirical findings.

and desires, as attitudes, are different because they have different functions in a system, that is, they each have typical causes and effects, typical relations with inputs, outputs, and other mental states. What makes my belief that it will rain a belief, and not a desire for rain, are the typical causes and effects it has. It can be caused by my perception of dark clouds in the sky, and may, along with my desire not to get wet, cause me to take an umbrella before leaving the house, etc. My desire for rain, on the other hand, does not have the same possible causes and effects as my belief that it will rain. This desire can be caused, for example, by my desire to stay at home doing nothing, and probably will not cause me to take an umbrella before leaving the house. What both these mental states have in common is that both are relations that individuals have with tokens of the same type of mental symbol, a symbol which represents that there will be rain here in the near future.

About mental processes, CTM starts with the assumption that they are "causal sequences of tokenings of mental representations." (FODOR, 1987, p. 17). In other words, thinking involves the transformation of mental symbols. If I believe it is raining, and I think that if it is raining, I will not go to the beach, and from that I decide to cancel my plan to go to the beach, what happened, according to CTM, was a causal sequence of symbols or mental representations. Fodor believes that mental states that are causally related are usually also semantically related. There is, in his view, a similarity between causal sequences of thoughts and arguments, which is, according to him, "one of the most striking facts about the cognitive mind as commonsense belief/desire psychology conceives it." (FODOR, 1987, p. 13). Fodor's favorite example is that of Sherlock Holmes, whose causal sequences of thoughts typically provide arguments for the conclusions to which he arrives. This is one of the central characteristics of the mind that can be explained by the assumption that mental processes are computational processes. As Fodor says,

> There is a well-known and, in my opinion, completely convincing argument for viewing the implementation of psychological laws in this way [as being computational]: It is characteristic of the mental processes they govern that they tend to preserve semantic properties like truth. Roughly, if you start out with a true thought, and you proceed to do some thinking, it is very often the case that the thoughts that the thinking leads you to will also be true. This is, in my view, the most important fact we know about minds; no doubt it's why God bothered to give us any. A psychology that can't make sense of such facts as that mental processes are typically truth preserving is ipso facto dead in the water. (FODOR, 1994, pp. 8-9).

Thus, when it comes to explaining these truth-preserving mental processes, Fodor combines his representational theory of mind with a computational theory of the mind, proposing the thesis that mental processes are computational processes. This means that sequences of thoughts are transformations of symbols or mental representations that occur mechanically by virtue of their syntactic forms (and not their contents), and that follow certain rules or algorithms.

The comparison of the mind with a computer, the computational theory of the mind, serves mainly to explain mental processes such as reasoning, i.e. to explain how, when we reason, one thought can cause another while preserving semantic properties such as truth.[9] A computer can be understood as "a device which processes representations in a systematic way." (CRANE, 2003, p. 85 and p. 121). In other words, computers process information through the algorithmic transformation of symbols. The transformation of symbols by a computer is guided by their syntax (or by their formal properties), and not by what these symbols means. As Fodor says, "computations just *are* processes in which representations have their causal consequences in virtue of their form." (1980a, p. 241). Fodor takes the syntactic form of a symbol to be "one of its higher-order physical properties. To a metaphorical first approximation, we can think of the syntactic structure of a symbol as an abstract feature of its shape." (1987, p. 18).

As an ultimately physical property, a symbols' syntax or shape can have causal power. But the manipulated symbols also have semantic content, for after all they are symbols. Although transformations, or causal sequences of symbols, occur only by virtue of the syntactic forms of symbols, the computer is so constructed that semantic relations between symbols can be preserved in these transformations. A computer can be programmed so as to transform symbols following the rule of modus ponens, for example, which preserves truth. When it does so, it outputs a symbol Q from a symbol P and the symbols $P \rightarrow Q$, and this causal sequence of symbols occurs only by virtue of the syntactic forms of those symbols, and not their contents. Since the syntactic form is ultimately a physical property, the computer shows us how it is possible for something mechanical to process information in a coherent way.

---

[9] According to Aydede, "thinking is not proceeding from thoughts to thoughts in arbitrary fashion: thoughts that are causally connected are in some fashion semantically (rationally, epistemically) connected too. If this were not so, there would be little point in thinking—thinking couldn't serve any useful purpose. Call this general phenomenon, then, the *semantic coherence* of causally connected thought processes. LOTH [the language of thought hypothesis] is offered as a solution to this puzzle: how is thinking, conceived this way, physically possible?" (AYDEDE, 2015). Strictly speaking, though, it is CTM that is offered to explain this (and not LOTH). While CTM presupposes a language of thought, the language of thought is not in itself sufficient to explain the semantic coherence of reasoning.

So while it is mysterious to see how meanings can have causal powers, it is easier to see how symbols can (via their syntax). The semantic properties of intentional states are derived from the semantic content of the mental symbols or representations that constitute them. And we explain the causal powers of intentional mental states by saying that they have causal powers in virtue of the syntax of mental representations.

If beliefs are relations that individuals have with mental representations, which are, as the LOT hypothesis says, linguistic symbols, and mental processes are transformations of these symbols, then we can explain these transformations by saying that, just as in a computer, mental symbols cause other symbols by virtue of their syntactic forms (which are ultimately physical), while preserving content relations. To the extent that beliefs (and other propositional attitudes) are relations to symbols, and that mental processes are computations, we can explain from a naturalistic point of view how there can be causal sequences of intentional states, such as in reasoning, that maintain a semantic coherence. As Fodor says,

> Computers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties. So *if* the mind is a sort of computer, we begin to see how you can have a theory of mental processes that succeeds where – literally – all previous attempts had abjectly failed; a theory which explains how there could be nonarbitrary content relations among causally related thoughts. (FODOR, 1987, p. 19).
>
> I don't know of any other remotely serious proposal for a mechanism that would explain how the processes that implement psychological laws could reliably preserve truth. (FODOR, 1994, p. 09)

But what does it mean to say that the syntactic form of a mental state is a higher-order physical property? As I understand Fodor's suggestion, it simply means that the syntax of a symbol is implemented by a physical thing, though it can't be reduced to something physical. Crane makes some observations that might help us understand this issue. He presents the LOT hypothesis as saying that having a belief or a desire is to have a certain relation to a mental representation, which is a sentence, in a sense, "written" in the head. He also makes a useful distinction between the *vehicle* and the *medium* of thought. So "when a thinker has a belief or desire with the content *P*, there is a sentence (i.e. a representation with semantic and syntactic structure) that means *P* written in their heads. The vehicles of representation are linguistic, while the medium of representation is the neural structure of the brain." (2003, p. 140). As he exemplifies,

"whenever someone believes, say, that *prices are rising*, the vehicle of this thought is a sentence. And the medium in which this sentence is realised is the neural structure of the brain" (CRANE, 2003, p. 137). The idea then is that sentences can be physically realized in different media: not only on paper, or in sound waves or electrical circuits of a computer, but also in the brain. Just as computers can store and transform sentences in the form of sequences of zeros and ones, which in the end correspond to electrical circuits, so we can suppose that something similar occurs in the brain, that is, that when we think, there are transformations of sentences of a mental language (which have semantics and syntax) which are carried out in the brain. The syntax of a symbol is taken to be a higher-order physical property to the extent that it is a formal feature of symbols that determines its causal powers, and that it is realized in a physical medium such as the brain, though it can't be reduced to it, for presumably the same symbol can be realized in different media.[10]

Thus, what CTM says is that thoughts have as their vehicle symbols or mental sentences, which are, at least in biological organisms, realized in the brain, and that the combinations and transformations of symbols occur by virtue of their forms, following certain transformation rules. The usual way to understand it is to think of the brain as hardware, and of the mind as a program that it instantiates.[11] Ultimately, what CTM says is that there is a legitimate level of explanation about the functioning of the mind that refers to symbols and their transformations, just as in a computer we can talk of the programming, which is a description which falls in between the fully physical level of the electrical circuits, and the level of the outputs. And if symbol manipulation is not mysterious in a computer (at least not for the programmers), neither should it be in the mind and brain.

In sum, the computational theory of mind gives a naturalistic account of mental processes, reasoning in particular (mainly, as we shall see later, local and demonstrative inferences). Thinking ceases to be conceived as an entirely abstract and immaterial activity, whose causal

[10] I suppose Fodor doesn't want to identify the syntax of a mental symbol to any of the properties of its neural realizers because he wishes to preserve the possibility of the multiple realizability of intentional states (as we will see in Chapter 2, section 2). It is worth noting that Fodor's notion of syntax is not unproblematic. See Pessin (1995) for some difficulties in Fodor's notion of mental syntax. Also, sometimes Fodor seems to take the syntax of a symbol to be nothing more than its non-semantic properties, which is a sense of "syntax" that is distant from its use in linguistics, for instance.

[11] We will see in chapter 2 that Searle, with the famous Chinese room argument, attacks the idea that programs are sufficient for the existence of intentional states. It should be noted, however, that Fodor does not advocate the idea that programs are sufficient for intentionality. For him, as we shall see, intentionality depends on certain causal relationships between the organism and the world.

power would be difficult to explain. It is considered to be transformations of symbols, which are physically realized in the brain and as such can be causally efficacious. Fodor clearly thinks that mental states should be treated as material entities. As he notes, he is not interested in the ontological question of whether beliefs and desires can be material, because he thinks the answer is "'yes,' because whatever has causal powers is ipso facto material" (1987, p. x).[12] But still, not all material things in the world have intentionality, or are about other things. RTM and CTM are intended to show us how the causal powers of intentional states can be made a little less mysterious. The causal transformations between mental states are supposed to be explained primarily at the symbolic level, not at the neural level of implementation. It is the syntax of the symbol, being ultimately something physical, which determines the causal power of that symbol. As Fodor says, "the syntax of a symbol might determine the causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open." (1987, p. 18). Again, we know that computers transform symbols mechanically according to certain rules, which guarantee, for example, the preservation of truth by transformations that take into account only the syntactic forms of symbols. If we suppose that mental processes are computational processes, i.e., syntactic transformations of symbols, we can begin to understand how mental processes can preserve semantic relations, while still being material processes. Thus, the causal power of intentional states, and the fact that the passage from one thought to another tends to preserve semantic properties such as truth, are no longer entirely mysterious. Computers show us how symbols can cause other symbols and preserve semantic relations at the same time.[13]

The idea that mental processes are computational processes is not Fodor's invention. What Fodor notes in *The language of thought* is precisely that the psychological theories at the time worked under this assumption.[14] In addition to his arguments in defense of CTM, Fodor contributes to the theorizing about the mind by making explicit a central assumption behind psychological theories that take mental processes to be computational: that "computation presupposes a medium of computation: a representational system." (FODOR, 1975, p. 27). More specifically, these theories presuppose that the vehicle of these representations is linguistic: "the theory that mental processes are computations depends on the theory that mental representations

[12] A substance dualist would reject this, but I won't consider this view here.
[13] The comparison of the mind to a computer program will be further discussed in chapter 2.
[14] Besides, there are several medieval and modern thinkers that accept the existence of something like a language of thought (see PANACCIO, 2017; CHOMSKY, 2009). Fodor himself admits that what he is doing is to "resurrect the traditional notion that there is a 'language of thought.'" (1975, p. 33).

are sentence-like; in particular that mental representations have constituent structures." (FODOR, 2007). That is, mental processes can only be computational processes – which, for Fodor, is what psychological theories assume they are – if there are mental representations which are like symbols, in that they have formal and semantic properties. Computations only occur if there are symbols that are being computed. Since computations are syntactic processes, it is being assumed that the mind is at least in part constituted by symbols with syntactic structure. "This emphasis upon the syntactical character of thought suggests a view of cognitive processes in general – including, for example, perception, memory and learning – as occurring in a languagelike medium, a sort of 'language of thought'." (FODOR, 1994, p. 09). Thus, if cognitive scientists are committed to the idea that mental processes are computational, they will be at least implicitly committed to the hypothesis that there is a language of thought, i.e. that there are mental representations that have combinatorial syntax and semantics. To the extent that Fodor accepts the computational theory of mind, he accepts the idea that the vehicle of thought (or at least of propositional attitudes and reasoning) is a mental language.

There is no doubt that talk about representations is pervasive in cognitive science. It is important to note, however, that representations are not assumed for the sole purpose of vindicating propositional attitudes admitted by common sense psychology (which is what RTM aims at). Cognitive scientists posit all sorts of representations, which are supposed to explain several aspects of our mental lives, for instance, different stages of perception, memory, language production and comprehension, etc., some of which are not accessible to the individual. So there is no need to assume that, whenever representations are assumed in a scientific theory, they will always be possible objects of propositional attitudes. Some of these symbols are assumed to be sub-personal – in the sense of not being (at least not usually) accessible to the individual. What RTM and LOT say, conceived as vindications of common sense belief/desire psychology, is simply that when one believes (or desires, etc.) something, and that state causes others, a mental representation with a compositional content and structure is being tokened (in this case, possibly accessible to the individual, though perhaps not always). But it leaves open the possibility of structured representations being tokened when there is no common sense attitude involving those representations.

Fodor often argues that our best scientific theories are our best indicators of what there is.[15] When he wrote *The language of thought* (1975), he emphasized that a good reason to accept that there is a language of thought was the fact that the only available psychological models, insofar as they dealt with mental processes as computational ones, presupposed mental representations with a linguistic structure. As he puts it, "if our psychological theories commit us to a language of thought, we had better take the commitment seriously and find out what the language of thought is like." (FODOR, 1975, p. 52). This argument may have lost some of its strength in the 1980s, with the popularization of connectionist models of the mind that, unlike so-called classical models, do not accept linguistically structured mental symbols.[16] However, as we are going to see in chapter 3, Fodor argues that connectionist models, unlike LOT, don't explain some important properties of thought (namely productivity and systematicity). So he still maintains that CTM is our best available cognitive theory and therefore our best indicator of what there is. If Fodor is right, the success of RTM gives us at least one good reason to suppose that there is a language of thought.

We have seen so far how RTM and CTM can be taken as naturalistic explanations of at least part of our mental life, namely, propositional attitudes and reasoning. As we have seen, they involve the assumption that propositional attitudes are *functional* relations that we have with mental symbols. This suggests that Fodor is an advocate of a functionalist theory of the mind, which characterizes mental states in terms of their causal or functional roles (that is, in terms of what they typically cause and what typically causes them), and not, for example, in terms of behavioral dispositions or their neural substrate.[17] However, as I intend to show in the next section, there are different types of functionalism, and someone who is committed to RTM and CTM doesn't need to accept all of them.

---

[15] For instance: "the question of how the mind works is profoundly interesting, and the best psychology we have is ipso facto the best answer that is currently available." (FODOR, 1975, p. viii). "I think the best kind of ontological argument [for the existence of mental representations as symbols] is the kind I've just given: we need this construct to do our science." (FODOR, 1981c, p. 29). "(…) some version or other of RTM underlies practically all current psychological research on mentation, and our best science is ipso facto our best estimate of what there is and what it's made of." (FODOR, 1987, p. 17).

[16] It might be said that connectionist models allow, but do not require, symbols with a linguistic structure. However, as we will see in chapter 3, Fodor and Pylyshyn distinguish the classical model (of the language of thought) from connectionist models precisely by saying that only the former accepts linguistically structured symbols. Insofar as connectionist models allow symbols, they are nothing more than models of implementation of the classical cognitive model. This way of distinguishing the two models is, I believe, accepted in the literature.

[17] These alternative theories (i.e. behaviorism and the identity theory) will be explored in chapter 2.

2. Fodor and functionalism

It is not so simple to say what Fodor's opinion on functionalism is. Although he says that propositional attitudes are functional relations that we have with mental representations, in some places he formulates explicit criticisms against functionalism. In this part of the chapter, we will see, first, that he criticizes the machine state functionalism proposed by Putnam. Then we will see that he also opposes what he sometimes calls semantic functionalism, which is the view that belief types, such as the belief that men landed on the moon, that it is going to rain, etc., may be individuated by their causal or functional roles. We will see that the types of functionalism Fodor criticizes (i.e. machine state functionalism and semantic functionalism) are different and independent from the kind of functionalism he proposes, which is a functional approach only to attitudes such as beliefs and desires, not to full propositional attitudes which involve particular mental representations. This allows Fodor to say that beliefs are functional relations with mental symbols, while denying that the belief that *P* can be individuated by its functional role.

A functional approach is primarily one that defines, characterizes, or individuates something by its functions, or by its causal role. Functionalism, as a theory of mental states, says that mental states are individuated by their causal roles, that is, by their typical causes and effects. According to Fodor,

> The intuition that underlies functionalism is that what determines which kind a mental particular belongs to is its causal role in the mental life of the organism. Functional individuation is individuation in respect of aspects of causal role; for purposes of psychological theory construction, only its causes and effects are to count in determining which kind a mental particular belongs to. (FODOR, 1981c, p. 11).

Putnam, in "Psychological Predicates" (1967), advocates the thesis that mental states are functional states, which are characterized by their causal relations with inputs, outputs, and other mental states. More specifically, Putnam makes use of the notion of a probabilistic automaton, which is essentially a nondeterministic Turing machine whose operation is described by a table. The table for the probabilistic automaton (which is assumed to receive sensory inputs and to produce motor outputs) specifies the probabilities of it passing from one state to another, or from

a state to a motor output, given certain inputs. A state in a machine table is individuated by the probabilities that other states attribute to it, and by the probabilities it attributes to other states. According to Putnam, we should conceive mental states as analogous to the states described in a machine table. Pain, for example, is a mental state that can be identified with a state described in a machine table, which is individuated by the probability that it has to follow certain inputs, and by the different probabilities it attributes to different outputs and other states of the machine. According to Putnam, "being capable of feeling pain *is* possessing an appropriate kind of Functional Organization" (1967, p. 163). What makes a mental state the state that it is is simply the way it fits into the organization of the system, that is, the probabilities it has to produce certain effects, and to be produced by certain causes. This position is usually called machine state functionalism.

Putnam, as we shall see in more detail in chapter 2, is primarily interested in countering behaviorism, as well as the view that mental states are identical to brain states (also known as the identity theory). In Putnam's view, behaviorism has the problem of not taking mental causation seriously, whereas the view that mental states are identical with brain states is too restrictive about the possible physical bases of a system with mental states. Identifying a mental state type with a state in a table, which specifies how this state relates probabilistically to other states, with inputs and with outputs, avoids the problems that the competing theories have.

Fodor is sympathetic to Putnam's criticisms to behaviorism and the identity theory, but he finds Putnam's functionalism too austere. In his paper "What psychological states are not" (1972), published with Ned Block, they reject the machine state functionalism proposed by Putnam. In the case of a table description of a probabilistic automaton, for one state to be identical to another, both need to have the same connections and probabilities assigned to them. Fodor and Block then argue that types of psychological states cannot be individuated in the same way. According to them, if we were to identify states in a machine table with types of mental states, we would have to say that people almost never share mental states.[18] They note that it seems appropriate to say that two people share the same pain when they stub a toe even if one is more likely to say "damn" and the other "darn". But if we accept machine state functionalism, we would have to say they are tokening different types of psychological states, given that the state

[18] As they put it, "on the assumption that there is a computational path from every state to every other, any two automata that have less than all their states in common will have none of their states in common." (FODOR; BLOCK, 1972, p. 93).

each one is in assigns different probabilities to possible outputs. That is, however, counterintuitive. Fodor and Block say, then, that "the conditions used to type-identify machine table states per se cannot be used to type-identify psychological states." (FODOR; BLOCK, 1972, p. 94). This form of functionalism individuates mental state types in a too fine-grained way.

Their own suggestion is that, in general, what serves to individuate types of entities are the scientific laws that subsume them. To identify fundamental physical entities we resort to the laws of physics that apply to them. Likewise, conditions for the individuation of psychological types will be provided by the psychological laws that apply to those states. Basically, psychological states are whatever true psychological theories and laws tell us they are.[19]

In short, Block and Fodor reject machine state functionalism, which says that mental states can be identified with states in a table describing a probabilistic Turing machine. But they do not explicitly reject the more general functionalist idea that mental states can be individuated at least in part by their functions; they only oppose the idea that what makes a mental state be the mental state that it is has anything to do with states in a machine table. They do indeed see advantages in functionalism as compared to competing theories, such as behaviorism and the identity theory. One is precisely that functionalism preserves the possibility that creatures with a non-biological constitution may have mental states (more on this in the next chapter).

In later writings, such as "The mind-body problem" (1981) and the introduction of *Representations* (1981), Fodor grants that there are some advantages in treating psychological states in terms of states that are described in a Turing machine table. One is that Turing machine programs are easily implementable mechanically. If mental states can be understood in terms of machine states, then, given that machine states can be mechanically realized, we can assume that mental states could also be implemented by a simple mechanism that manipulates symbols in a purely syntactic way, without presupposing any intelligence, or any homunculus.[20] But Fodor still thinks that "what Turing machine functionalism provides is just a *sufficient* condition for the mechanical realizability of a functional theory." (1981c, p. 14). He also says, alluding to his paper

[19] It is unclear that this would work for specific mental states. Though psychology may have something to say about beliefs, desires, fears, hopes, etc., there probably will not be psychological laws about, say, the desire to eat a white Kit Kat.

[20] According to Fodor, "a psychological explanation that can be formulated as a program for a Turing machine is ipso facto mechanically realizable even though, *qua* functional explanation, it is not a specification of a mechanism." (1981c, p. 14).

with Block, that "there are reasons for supposing that the relativization to Turing machines is much too austere for a psychologist's purposes" (idem, p. 14).

Another problem with the functional characterization of mental states which Fodor and Block note is that functionalism does not seem to offer a satisfactory characterization of mental states with qualitative content, such as pain and itching. Functional individuation leaves out the qualitative aspect of sensations, which seems to be essential to them.[21] Aside from being caused by, say, damage to the skin, and being likely to cause screaming, a pain *feels* a certain way. Functionalism allows for the possibility that a state that has the same causal role as pain, but which does not feel like pain, still counts as pain. Intuitively, this doesn't seem right.

However, Fodor does seem to believe that propositional attitudes such as beliefs and desires, which presumably have no characteristic qualitative content, can be characterized, at least partially, in a functional way – though perhaps not exactly in terms of a Turing machine table. But, according to him, the functionalist view alone does not give us a complete characterization of these mental states. Beliefs and desires are states with semantic content, and to explain how these states can be intentional, it is not enough to consider them as functional states. Functionalism alone seems compatible with the view that beliefs and desires are not semantically evaluable, which is contrary to a central assumption of common sense psychology that we wish to preserve. So if we want a way of conceiving these states which doesn't leave their intentionality behind, we should add to their characterization the notion of mental representation. That is, we should say that propositional attitudes are functional relations *with mental representations*, which is where their intentionality comes from. It is precisely here where the representational theory of the mind enters the picture, adding something to functionalism.

Let us now turn to Fodor's critique of what he calls semantic functionalism. There are two possible functionalist approaches to propositional attitudes. A first would be to say that attitudes such as believing and desiring are functional states, and a second would be to say that believing in *P* or desiring *Q* are functional states. In his early writings, Fodor does not draw this distinction. In "The mind-body problem", for example, he seems to accept these two forms of functionalism. However, in later writings, such as *Psychosemantics* (1987), Fodor strongly criticizes the second type

[21] Kripke raises a similar objection to the identity theory and to functionalism: "this notion seems to me self-evidently absurd. It amounts to the view that *the very pain I now have* could have existed without being a mental state at all." (1980, p. 147).

of functionalism, which is the idea that belief and desire types, such as the belief that it is going to rain, or the desire to eat chocolate, may be individuated by their causal roles. This kind of functionalism, which proposes that a belief together with its semantic content can be individuated by its causal role is what Fodor calls semantic functionalism, and he rejects it. The kind of functionalism he advocates is the first, according to which attitudes such as believing and desiring, taken in general, can be individuated by their causal roles:

> I suppose we like Psychofunctionalism (those of us who do) because it provides us with a reasonable account of the difference between systems to which belief/desire ascriptions are appropriate and systems to which they are not. "It's a question of internal organization," Psychofunctionalism says. "People and machines have the right sort of internal organization to satisfy belief/desire explanations; rocks and such, by contrast, don't." (...) If, however, *that's* what you want Psychofunctionalism for, then all you need is the claim that *being a belief* is a matter of having the right connections to inputs, outputs, and other mental states. What you *don't* need – and what the philosophical motivations for Psychofunctionalism therefore do not underwrite – is the much stronger claim that being the belief *that P*, being a belief that has a certain *content*, is a matter of having the right connections to inputs, outputs, and other mental states. (FODOR, 1987, p. 69).

Fodor thinks, then, that beliefs and desires, considered as attitudes (leaving aside their particular content), can be characterized in a functional way. Beliefs have typical relations with inputs, outputs, and other mental states, whereas desires, hunches, intentions, etc., have different typical relations. What Fodor rejects is *semantic* functionalism, which is the view that the belief or desire that *P* can be individuated in functional terms. This kind of functionalism is called semantic functionalism because it is a functionalism that applies to the semantic contents of intentional states. The idea, then, according to this kind of functionalism, is that what makes the belief that it is going to rain a belief with that specific content is the causal role it plays within a system. According to this view, the difference between believing *P* and believing *Q* is a difference in the causal roles of each belief.

Fodor argues, however, that the belief that *P* (that it is going to rain, for example) is not a type that can be identified with its causal relations with other mental states, inputs, and outputs.[22]

[22] As he says in *Psychosemantics*, "philosophers of 'functionalist' persuasion even hold that the causal powers of a mental state determine its identity (that for a mental state to be, as it might be, the state of believing that Demetrius killed Lysander is just for it to have a characteristic galaxy of potential and actual causal relations). This is a position of some interest to us, since if it is true – and if it is also true that propositional attitudes have their contents essentially

The problems with semantic functionalism are essentially the same as with machine state functionalism. The central problem, according to Fodor, is that semantic functionalism implies meaning holism (the idea that the meaning of a mental representation is determined by the meaning of all other mental representations to which it is connected)[23], which in turn implies the falsity of intentional realism (the idea that it is possible for people to share the same intentional state, and that we can form generalizations about them). If the belief that *P* is individuated by its causal relations to all other mental states, and if the belief that *P* has different causal roles in different people, it can't be true that different individuals, or even the same individual at different times, share the same belief.[24] This would make it a miracle that our daily psychological explanations seem to work, and remember that Fodor is primarily interested in securing common-sense psychology. Fodor wants to vindicate our everyday explanations of behavior, and these explanations seem to be based on the assumption that different people share beliefs and desires.

Let us consider an example: the desire to eat chocolate. It does not seem right to say that there are typical inputs that lead to this desire, nor typical responses to it. As far as the inputs of this desire are concerned, imagine that I, for example, always feel like eating chocolate when I realize it's noon. Obviously, this is an idiosyncrasy of mine, and it doesn't seem appropriate to characterize the belief that it is noon as a typical input that leads to the desire to eat chocolate. The same state can be a typical cause of the desire to eat chocolates for one person, not another. John may want to eat chocolate when he sees chocolates, while seeing that there is chocolate around may cause aversion in Mary. This kind of variation also holds true for the outputs of this desire. Nothing guarantees that the desire to eat chocolate will lead one to eat chocolate, or even to try to eat chocolate. One can eat chocolate, but one can equally well decide to eat an apple instead, or to do a headstand. Each person, or the same person at different times, may have different causal roles associated with the desire to eat chocolate. Since there are no unique causal

– it follows that the causal powers of a mental state somehow determine its content. I do not, however, believe that it is true." (FODOR, 1987, p. 12).

[23] Fodor's argument for this view is that there is no non-arbitrary way of determining which inferential or causal relations are the ones that serve to individuate a mental state. Or, as he puts it, "you can't have a principled distinction between the kinds of causal relations among mental states that determine content and the kind of causal relations among mental states that don't." (FODOR, 1990, p. x). This is why semantic functionalism leads to holism.

[24] Fodor sometimes calls this view conceptual (or inferential) role semantics. According to him, "all such theories are inescapably infected with holism and are therefore incompatible with the working assumption that the laws of psychology are intentional. If what you're thinking depends on *all* of what you believe, than nobody ever thinks the same thing twice, and no intentional laws ever get satisfied more than once; which is tantamount to saying that there aren't such laws." (1994, p. 06).

roles that are always present when this desire occurs, it does not seem appropriate to individuate it by its causal roles. If we were to identify it by some of its causal roles, we would certainly have to choose one arbitrarily. This would force us to conclude, in situations where we would normally say that someone wants to eat chocolate, but in which the desire in question does not have the same arbitrarily chosen causal role, that the person in fact does not have the desire to eat chocolate. If we accepted semantic functionalism, the intuitive idea that we often share the same beliefs and desires with other people, even when those states have different causal roles in each of us, would be false.[25]

In the end, Fodor's main criticism of semantic functionalism is very similar to his central objection to taking pain to be a state in a probabilistic machine. If we accepted Putnam's machine state functionalism, we would have to conclude that different people rarely, if ever, share mental states like pain, because they rarely attribute the same probability to its possible outputs. Likewise, if particular beliefs were individuated by their causal roles, assuming that beliefs have different causal roles in different people, we would have to accept that different people (and the same person at different times) hardly share the same beliefs.

Fodor therefore rejects both machine state functionalism and semantic functionalism, and accepts a weaker functionalism that applies to propositional attitudes, to which Fodor adds RTM and CTM. In accepting a weaker functionalism, Fodor avoids problems associated with the stronger forms, while still preserving a feature he finds desirable in functionalism: functional explanations given by psychology are autonomous from the explanations of lower-level sciences, such as neuroscience (at least when it comes to propositional attitudes). As he notes,

> Functionalism just *is* the doctrine that the psychologist's theoretical taxonomy doesn't need to look "natural" from the point of view of any lower-level science. This seems to some neuroscientists, and to some of their philosopher friends, like letting psychologists get away with murder. (FODOR, 1990, p. 10).

[25] A proponent of semantic functionalism could say that, e.g., the desire to eat chocolate is defined by all its *possible* relations to inputs, outputs and other mental states. So even though the actual inputs and responses of the desire to eat chocolate may vary from person to person, and for the same person from time to time, still the possible inputs and outputs are shared. I think, though, that this way of type-individuating mental states would be very little informative about what tokens of, say, the desire to eat chocolate, in fact have in common, so as to be taken as occurrences of the same desire. Besides, one can imagine that two states share all possible input-output connections, while still intuitively being different states.

RTM and CTM, as we have seen, characterize propositional attitudes as functional relations that we have with mental symbols, and mental processes as computational transformations of symbols. Nothing is said, therefore, about the material which makes up a system capable of having beliefs and desires, or one capable of thinking. This system should be organized in such a way as to make it capable of processing information in a relevant way, but the means by which this information processing is achieved is not particularly relevant for psychology. This sort of functionalism, which Block calls "computation-representation functionalism", takes psychological explanation to be a kind of functional explanation,

> seen as akin to providing a computer program for the mind. Whatever mystery our mental life may initially seem to have is dissolved by functional analysis of mental processes to the point where they are seen to be composed of computations as mechanical as the primitive operations of a digital computer – processes so stupid that appealing to them in psychological explanations involves no hint of question-begging. (BLOCK, 1980, p. 183).

Insofar as the functioning of an organism is explained by specifying the computations it performs, according to this conception, in order to describe the psychological functioning of an organism it is not necessary to describe the functioning of its brain (it may not even have a brain). Psychological explanations are therefore independent from neurological explanations (we will return to this in the next chapter). Thus, although Fodor doubts that the individuation of mental states can be fully functional, he certainly accepts that the kind of explanation to be offered by psychology is in some sense functional, if computational explanations can be taken as functional.

Again, what Fodor thinks can be characterized in a functional way are attitudes like believing and desiring, each having typical causal roles. One problem with this is that, to my knowledge, Fodor does not give substantial characterizations of the typical causal roles of each attitude. In some places he briefly suggests that perception is a typical cause of beliefs, and that intentions usually cause actions that have the purpose of obtaining whatever one had the intention to obtain.[26] But these characterizations, it seems to me, are too vague. Perception is certainly a common cause of beliefs, but it also seems to be a common cause of desires, of

[26] As Fodor says, "seeing that a is F is a normal cause of believing that a is F; the intention that it should be the case that so and so is a normal cause of actions whose goal is to bring it about that so and so (…)" (1981c, p. 25).

expectations, of hunches, etc. And it seems that beliefs, desires, fears, etc., can all be the cause of actions.

One way out of this problem would be to say, with Crane, that the functional differences between attitudes is something to be investigated by cognitive scientists. According to him (CRANE, 2003, p. 141), it is not the goal of the proponent of LOT to explain the difference between beliefs and desires, but rather to explain what it is to have an attitude with one content rather than another. But still, not providing a precise characterization of the attitudes is, I think, a downside of Fodor's RTM.

3. Scope and limitations of RTM and CTM

> "saying the last word isn't the problem that currently confronts us. Our problem is to find something—almost anything—that we can say that's *true*."
> (Fodor, 1975)

Before proceeding, it is important to have some idea of the scope and limitations of RTM and CTM. Although Fodor sometimes says in books such as *Psychosemantics* and *Theory of Content* that they vindicate common sense psychology, we must insert some qualifications into this observation. Also, although Fodor may sometimes seem optimistic about the scope of CTM, in *The language of thought* he already thought that CTM most likely would not be able to explain a number of mental processes. His negative remarks about the scope of CTM reappear in more recent books such as *The mind doesn't work that way* and *LOT 2*. Fodor's idea is that although the computational theory of mind is the best available theory for rational mental processes, it most likely is not the whole truth about the workings of the mind. In what follows I will point out some limitations of RTM and CTM, in particular RTM's inability to explain conscious qualitative states, and CTM's inability to explain abductive mental processes.

3.1 Limited vindication of common sense psychology

Fodor often treats the expression "common sense psychology" as interchangeable with the expression "belief/desire psychology". As we have seen, when he proposes to vindicate common

sense psychology, he wants to provide a naturalistic theory that accepts entities with the properties that common sense attributes to mental states, namely, semantic evaluability and causal efficacy. However, it is important to keep in mind that common-sense psychology does not make use only of explanations of behavior that refer to semantically evaluable states, such as beliefs and desires. It is part of our daily psychological explanations to say that Julia went to sleep because she was tired, that Peter ate too much because he was sad and anxious, that Sophia put on a coat because she was cold and wanted to warm up, that Jack took a medicine because he was in pain and wanted it to stop, and so on. It is questionable that sensations and emotions, such as tiredness, sadness, anxiety, cold and pain can be treated as intentional states (or that being semantically evaluable by their relation to the world is one of their essential characteristics).

I'm not denying that sensations and emotions may represent something in the world. It may well be that, for some of these states, representing something is one of their essential features. All I'm claiming is that, if anything is essential to them, it is their qualitative character. In normal situations, I can say that my sensation of heat represents, approximately, what the temperature in the world is. This representation may be accurate or not – I may have a fever, for instance, in which case a representationalist could say that what my sensation of heat represents is my internal condition of fever. But it is unclear that the same treatment could be given to all sensations – to pain, for instance. Even if we accept that, e.g., the sensation of heat is representational, representing something is unlikely to be its only essential feature. For if anything undoubtedly seems essential to them, it is that they have a certain qualitative content, or, to use Nagel's expression, that there is something that it is like to be in one of these states. Besides, even if they are partly representational states, it is not clear that they have the right format to take part in computational processes.

If we accept that not all sensations and emotions can be treated as representational states, then RTM and CTM would not be natural candidates to explain them. And for a naturalistic psychological theory to completely vindicate common-sense psychology, it would have to explain not only how beliefs and desires may have causal power, but also how sensations and emotions can cause and be caused by behaviors and other mental states. In short, a complete vindication of common sense psychology requires a naturalistic theory that explains not only causally efficacious representational states, but that explains also how emotions and sensations such as envy, love, heat, pain, etc. can have qualitative character and causal powers.

As we've seen, Fodor's solution to the problem of presenting a naturalistic construct of beliefs and desires, compatible with a materialist view of the mind, is to say that these states are relations we have with symbols, and that computers show us how these symbols can undergo causal transformations by virtue of their syntactic forms. But at least some sensations and emotions do not seem to be representational mental states, states which can be characterized solely as relations that we have with linguistically-structured symbols. And if they are not relations with symbols, then their causal power cannot be treated as analogous to the causal powers of symbols in computational processes; computers don't give us an appropriate mechanism to conceive them.

Fodor himself has conceded that for mental states like sensations, their phenomenological or qualitative character seems essential to them, and, as we've seen, he admits that functionalism does not offer a good theory about these states:

> It seems to me, for what it's worth, that functionalism does not provide an adequate account of qualitative content; that, in fact, we have no adequate materialist account of qualitative content. Since I have nothing at all to add to this discussion – either here or in the papers that follow – I shall simply assume that the functionalist program applies at most to the analysis of propositional attitudes. (FODOR, 1981c, p. 17).

Fodor, in general, says very little about states with qualitative content. Even in later works, as far as I know, he does not offer an explanation of these states, although he often says, as we have seen, that he intends to vindicate common-sense psychology. This may initially suggest that Fodor attributes a perhaps disproportionate importance to mental states such as beliefs and desires, and that he has a narrow view of what constitutes common sense psychology. In his defense, however, I do not think Fodor would deny that mental states like emotions and sensations are an important part of common sense psychological explanations of behavior (though he in fact never refers to these states when he talks about common sense psychology). What happens is that, as Fodor concedes in the above passage, he does not know how to explain qualitative states in a naturalistic way. So when Fodor says that his theory vindicates common sense psychology, by "common sense psychology" Fodor only means "belief/desire (or intentional) common sense psychology". Sometimes he is explicit about this, but sometimes he is not. So it is not that he thinks that common sense explanations of behavior only appeal to

intentional states, it is just that these are the only states he has some idea how to explain. So contrary to what may initially have appeared, Fodor does not claim to offer a full vindication of common-sense psychology because he does not know how to formulate a naturalistic psychological theory of qualitative mental states.

But still, whatever Fodor's intentions were, one could still complain that his theory is not sufficiently comprehensive in that it proposes to explain only a limited part of our mental life, which comprises only propositional attitudes and mental processes which are driven solely by the syntactic form of mental symbols. If it is true that qualitative mental states are not representational, then RTM and CTM will have nothing to say about them. There will be no part for a language of thought to play in their explanation. And assuming that qualitative states prove to be at lest in part representational states, it seems unlikely that RTM and CTM could tell us all there is to tell about their nature.[27] So if what we are after is a unified theory that addresses both intentional and qualitative states, RTM and CTM may not be the better place to look. This means then that the language of thought cannot be regarded as a hypothesis that explains the nature of all mental states.

One way out of this objection would be to say that there is no reason to expect that a unified theory can be provided. In addition, explaining anything about the mind, even if only propositional attitudes and processes involving them, is already better than nothing. In *The language of thought* Fodor thinks that it may well be that conscious states (which he seems to take as synonymous of qualitative states) can only be given a biological, not a cognitive explanation. In fact, as he admits "I try never to think about consciousness. Or even to write about it." (1998b, p. 73). In *The language of thought*, he seems to assume that an explanation at the cognitive level necessarily makes reference to symbols, or to a language of thought. If qualitative mental states cannot be explained in these terms, they will be explained only by a lower-level science, such as neuroscience. In Fodor's defense, though, not explaining conscious/qualitative states[28] is a

[27] Though it is usual in cognitive psychology to treat the different perceptual systems, for instance visual perception, as involving inferential processes. In Fodor's view, this means that "perceptual processes involve computing a series of redescriptions of impinging environmental stimuli. But this is to acknowledge that perception presupposes a representational system." (FODOR, 1975, p. 51). So it can be that some states that have a qualitative character, like states involved in visual perception, involve symbolic manipulation, even though symbolic manipulation may not be sufficient to explain their qualitative character.

[28] I think it is reasonable not to take conscious states to be coextensive with qualitative states. Propositional attitudes, for example, don't seem to have characteristic qualitative features, though they do seem to be either conscious or unconscious. So another possible objection to RTM is that it provides no way of distinguishing between conscious and unconscious beliefs and desires, which means that it is not offering a complete vindication to even the fraction of

problem that affects most theories of mind, so it is not a particularly strong objection to his theory. Assuming that RTM does give a proper characterization of propositional attitudes, demanding it to also handle sensations and emotions (as well as the qualitative character of perceptual states) would be asking too much of it.

3.2 Limited view of mental processes

RTM and CTM thus seem to have a restricted scope in that they explain only propositional attitudes, and mental processes involving them. But initially it could be said that while they only explain mental processes involving representational states, at least that part of human psychology they explain entirely. However, since *The Language of Thought*, Fodor already recognized that not all types of mental processes, not even all of those involving representational states, could be treated as computational transformations of representations. There he said that perhaps the most that a computational theory could explain were rational mental processes. According to him, "a theory of the structure of rationality is the best that we can hope for with the tools we have in hand." (FODOR, 1975, p. 203). In later books, he still thinks that CTM explains rational mental processes, "the frequent similarity between trains of thought and *arguments*." (1987, p. 13). It explains, above all, how rational mental states can occur mechanically.[29] According to the computational view of the mind, thoughts cause other thoughts by virtue of their syntactic forms, and not their semantic contents, but relations of content are preserved in these transformations. Arguments that follow the rule of modus ponens, for example, show us how semantic properties, such as truth, can be preserved in the conclusion only by virtue of the forms of the premises. What Turing shows, according to Fodor, is that this kind of transformation can easily be implemented mechanically. Computers can be programmed so as to transform symbols in a mechanical way, which is sensitive only to the their syntactic forms, but

---

common sense psychology that it is supposed to vindicate. But here again, we could expect this to be explained by a different theory. Or, it could be left to psychology to show that conscious and unconscious propositional attitudes have different causal roles.

[29] According to Fodor, "to miss the point of going on about syntax/semantics parallelism is to miss what seems to me to be the main philosophical interest of the computational approach to psychology. There are, I think, three great metaphysical puzzles about the mind. How could anything material have conscious states? How could anything material have semantical properties? How could anything material be rational? (where this means something like: how could the state transitions of a physical system preserve semantical properties?) The parallelism story – if, indeed, it can be made to work – answers question three." (1991b, p. 285).

which is not arbitrary, because they can be made to follow, for example, truth-preserving rules. Computers, when programmed to make these transformations according to logical rules, show us how some forms of reasoning can be handled mechanically. If mental processes are often like arguments, we can explain them in terms of computational processes, which means that we can see how something material can be rational.

As I mentioned earlier, in *Psychosemantics* and *Theory of content*, Fodor gives us the example of Sherlock Holmes, whose mental processes often resemble arguments. Fodor goes so far as to say that "Conan Doyle was a far deeper psychologist – far closer to what is essential about the mental life – than, say, James Joyce (or William James, for that matter)." (1990, p. 21). By this, Fodor means that it is more essential to the mental life that our reasonings often take the form of arguments than that mental processes are often associative (a popular view among empiricists, like Hume). The computational theory of mind, according to Fodor in these books, would then be able to explain the most essential kind of mental process: the passages from one thought to another that resemble arguments.

We can hence draw the *first* limitation of the theory regarding its treatment of mental processes: associative mental processes are excluded from its scope.[30] From Fodor's point of view, the computational theory of mind does not explain a quite common kind of mental process, which is the passage from one thought to another by at least seemingly arbitrary associations, without one thought causing another in a way which respects rational relations between their contents. According to Fodor,

> cognitive psychology is about how rationality is structured, viz., how mental states are contingent on each other. (…) Cognitive explanation requires not only causally interrelated mental states, but also mental states whose causal relations respect the semantic relations that hold between formulae in the internal representational system. The present point is that there may well be mental states whose etiology is precluded from cognitive explanation because they are related to their causes in ways that satisfy the first condition but do not satisfy the second. (FODOR, 1975, p. 202).

Fodor gives the following example of an associative mental process, which would not be explained by CTM: in order to be reminded later to send a message to a friend, one can adopt

[30] At least in Fodor's conception of CTM. For criticisms to his views about the limited scope of CTM, see Pinker (2005).

the strategy of putting his watch on upside down. So when he goes check the time, he sees the watch upside down and is reminded that he needs to send a message to his friend. What occurs there is an associative causal sequence of mental states (seeing the watch causes him to remember that he needs to send a message to a friend), in which the content of the first state has no rational relation to the content of the second state. In *The language of thought*, Fodor sees it as a limitation of CTM that it cannot offer an explanation of associative mental processes like this.[31] According to him,

> I think it's likely that there are quite a lot of kinds of examples of causal-but-noncomputational relations between mental states. Many associative processes are probably like this, as are perhaps, many of the effects of emotion upon perception and belief. If this hunch is right, then these are bona fide examples of causal relations between mental states which, nevertheless, fall outside the domain of (cognitive) psychological explanation. What the cognitive psychologist *can* do, of course, is to specify the states that are so related and say *that* they are so related. But, from the psychological point of view, the existence of such relations is simply a matter of brute fact; explaining them is left to lower-level (probably biological) investigation. (FODOR, 1975, p. 203).

That is, Fodor basically thinks that associative processes probably cannot be treated as computational processes. As Fodor conceives the research program in cognitive psychology, it involves dealing with causal sequences of mental states as computational transformations of mental representations, where the contents of these representations are related. If we accept that associative processes cannot be approached in this way, presumably because there is in them no parallelism between syntax and semantics, then they cannot be the object of study of cognitive psychology. As Fodor says, the theory advocated in the book

> requires of a mental state that it be analyzable as a relation to a representation, and that its causal antecedents (or consequents or both) should be analyzable as relations to semantically related representations. This is, I think, a condition on a *rational* relation between events in a mental life, and I suppose that it's a point of definition that only relations which are in this loose sense rational can have a chance of being analyzed as computational. (FODOR, 1975, p. 203).

[31] Fodor is also skeptical about the prospects of the theory when it comes to explaining creative processes, like the writing of a poem, or the discovery of a scientific law (cf. FODOR, 1975, p. 201).

In later writings, Fodor directs several criticisms against the associationism typical of modern empiricists, and against connectionist models of the mind, which he considers to be a manifestation of an associationist trend in the cognitive sciences. One of his arguments against connectionism is precisely that it is only based on associations, and therefore does not explain the rational passage from one mental state to another.[32] Crane notes that although Fodor acknowledges that we can sometimes think through associations, he thinks this cannot be the main type of mental process that we use. As Crane argues, if we always thought associatively, we could hardly explain people's behavior, since each would presumably make their own random associations of thoughts. Instead, we usually attribute rationality to others, and this helps us explain and predict their behavior (cf. Crane, p. 143). But even if this is true, it seems undeniable that we do, at least sometimes, think by associations. And not explaining associative thought processes ends up limiting the scope of CTM to rational mental processes.

The problem is that, even restricted to rational processes, it is questionable that CTM, as Fodor conceives it, can explain all of them. So a *second* limitation of the computational theory seems to be that it doesn't really explain certain non-demonstrative inferences, such as abductions or inferences to the best explanation.[33] Fodor obviously acknowledges that we do not always reason in a demonstrative way. In fact, he initially seemed to believe that CTM was a good model for explaining rational mental processes in general (demonstrative or otherwise, where the conclusion preserves truth, or simply plausibility, warrant, etc.). In his paper "Connectionism and cognitive architecture" (1988), published with Zenon Pylyshyn, they are optimistic about using CTM to explain non-demonstrative mental processes. They say they deal first with valid reasoning because its formalization is the most well-known, but they expect that other kinds of inference can also be treated syntactically:

> It needn't, however, be strict truth-preservation that makes the syntactic approach relevant to cognition. Other semantic properties might be preserved under syntactic transformation in the course of mental processing – e.g., warrant, plausibility, heuristic value, or simply semantic non-arbitrariness. The point of Classical modeling isn't to characterize human thought as supremely logical; rather, it's to show how a family of types of semantically coherent (or knowledge-dependent) reasoning are mechanically possible. Valid inference is

[32] As Fodor says, "exactly what was wrong with Associationism, for example, was that there proved to be no way to get a rational mental life to emerge from the sorts of causal relations among thoughts that the 'laws of association' recognized." (1987, p. 18).

[33] As least when we consider non-modular, domain general thoughts.

> the paradigm only in that it is the best understood member of this family; the one for which syntactical analogues for semantical relations have been most systematically elaborated. (FODOR; PYLYSHYN, 1988, p. 29, fn).
>
> Classical theory construction rests on the hope that syntactic analogues can be constructed for nondemonstrative inferences (or informal, common-sense reasoning) in something like the way that proof theory has provided syntactic analogues for validity. (idem, p. 30).

Until the 2000s, Fodor holds that reasoning (whether demonstrative or not) is the most essential feature of our minds, and that this is the central class of mental processes that his computational theory is intended to explain. However, in more recent books, such as *The mind doesn't work that way*, Fodor highlights stronger limitations to the computational theory of mind, recognizing that it does in fact appear to be incapable of explaining rational, non-demonstrative mental processes such as abductive reasoning. There, Fodor casts doubt on cognitive scientists, such as Steven Pinker, who, according to him, believe that the mind can be fully explained computationally. Fodor's central point is that computational processes, as classical CTM understands them, are sensitive only to the intrinsic local properties of mental representations, such as their syntactic form (the parts of a mental representation and the way they are combined). Demonstrative reasoning forms a class of mental processes that can be described as transformations of mental representations that take into account only their local syntactic properties. An inference from A & B to A does not take into account anything that is outside the first representation, and therefore can be said to be local. Abductive inferences, on the other hand, take into account global, contextual properties of the system of representations (or of beliefs) of which the inferred representation is part. When we acquire a belief by an abductive inference, we presumably take into account its simplicity, or its relevance and coherence. But Fodor points out that the simplicity of a belief, or its relevance and coherence, are extrinsic properties, which depend on the context, or the belief system in which it is embedded. That is, simplicity, relevance and coherence seem to be examples of properties that are globally determined. And, according to Fodor, it is questionable that global properties can be plausibly reduced to local syntactic relations[34]:

[34] Fodor accepts that some mental processes are nondemonstrative and local: "I don't, of course, claim that processes of nondemonstrative inference are never local. To the contrary, devising and confirming theories of local inferential cognitive processes is exactly what cognitive science has proved to be a success at." (2008, p. 114, fn. 24). I suspect

> I think it's likely that a lot of rational belief formation turns on what philosophers call "inferences to the best explanation." You're given what perception presents to you as currently the fact, and you're given what memory presents to you as the beliefs that you've formed till now. Your cognitive problem is to find and adopt whatever new beliefs are best confirmed on balance. "Best confirmed beliefs on balance" means something like: the strongest and simplest relevant beliefs that are consistent with as many of one's prior epistemic commitments as possible. But, as far as anybody knows, relevance, strength, simplicity, centrality, and the like are properties, not of single sentences, but of whole belief systems; and there's no reason at all to suppose that such global properties of belief systems are syntactic. (FODOR, 1998c, pp. 205-6).

> Computation is, by stipulation, a process that's sensitive to the syntax of representations and to nothing else. But there are parameters of beliefs (and hence, RTM being assumed, of the mental representations that express them) that determine their role in nondemonstrative inference but are, on the face of them, not syntactic: relevance, conservatism, simplicity are extremely plausible examples. So, either learning, perception, belief fixation, and the like aren't processes of nondemonstrative inference (but what on earth else could they be?) or they aren't computations. (FODOR, 2008, p. 124).

Fodor acknowledges in *The mind doesn't work that way* that a property like simplicity, although global, could still be, in theory, reduced to the syntactic properties of the representational system taken as a whole. That is, simplicity is not a local, or intrinsic, syntactic property of a representation, but in principle it could be treated as a relational syntactic property, one that a representation maintains to a belief system. The problem, Fodor points out, is that computational processes, as the classical CTM understands them, are always sensitive only to the local properties of representations. It would only be possible to treat simplicity in a purely syntactic way if the whole belief system was taken as the minimum unit upon which the computations operate. But this approach, besides being impractical from a computational point of view, does not seem plausible from a psychological point of view. When we adopt a new belief, we do not seem to confirm it by taking as our basic unit of evaluation our entire belief system. A

---

that Fodor is here thinking of modular processes, such as those that occur in early stages of visual perception, which are assumed to be encapsulated (they have no access to information outside the module). As he says in *The modularity of mind*, "*any* nondemonstrative inference can be viewed as the projection and confirmation of a hypothesis, and I take it that perceptual inferences must in general be nondemonstrative, since their underdetermination by sensory data is not in serious dispute." (FODOR, 1983, p. 69). What is important to note is that, as much as there may be non-demonstrative mental processes that are local (the ones that are encapsulated), and that therefore can be treated computationally, there are pervasive rational mental processes, such as abductions, which are global, domain general (unencapsulated), and which Fodor believes cannot be adequately addressed by CTM.

computational explanation of global processes, in Fodor's view, can only work "at the price of a ruinous holism", which is not a plausible model of how the mind actually operates in such cases.

> it's *just got to be* possible to determine, with reasonable accuracy, the impact of adopting a new belief on one's prior epistemic commitments without having to survey those commitments in their totality. *Whole theories* can't be the units [of] computation any more than they can be the units of confirmation, or of assertion, or of semantic evaluation. The totality of one's epistemic commitments is *vastly* too large a space to have to search if all one's trying to do is figure out whether, since there are clouds, it would be wise to carry an umbrella. (FODOR, 2001b, p. 31).

It is not plausible to suppose that our entire belief system is the basic unit of confirmation or computation each time we acquire a new belief in a nondemonstrative way. We somehow select what is or is not relevant in order to adopt or not a new belief. But the problem is that, in principle, any previous belief may be relevant to the adoption of a new belief, and it does not seem possible to demarcate a priori which beliefs are relevant. As Fodor says,

> Reliable abduction may require, in the limit, that the whole background of epistemic commitments be somehow brought to bear in planning and belief fixation. But feasible abduction requires, in practice, that not more than a small subset of even the relevant background beliefs is actually consulted. How to make abductive inferences that are both reliable and feasible is what they call in AI the frame problem. (FODOR, 2001b, pp. 37-8).

It should be mentioned that Fodor is a proponent of the modularity thesis (1983), Roughly, according to it, the mind is in part constituted of modules, that is, cognitive mechanisms that operate on specialized and "encapsulated" domains of information. A module, as Fodor uses that notion, deals with a limited domain of information, and does not have access to information processed by other modules. In addition, a module's processing of information is fast and automatic. Examples of modules within the visual system are the "mechanisms for color perception, for the analysis of shape, and for the analysis of three-dimensional spatial relations." (FODOR, 1983, p. 47). Modular processes are good candidates for being computationally tractable because they deal with limited domains of information.

But in addition to modules, Fodor thinks there are also central processes of information, which are neither encapsulated nor specialized. This type of processing has access to the output of input (or perceptual) systems, and can be used, for example, to correct errors in a module's output. In order to do this, though, it needs to have access to information from different domains. Abductive inferences are also part of central information processing, since they do not deal with a specific or encapsulated domain of information. In theory, they can take into account any previous beliefs, regardless of what they are about. What Fodor basically thinks is that CTM can be successful in dealing with modular information processing, because such processes are computationally tractable. Central processes, though, are global and domain general, so they cannot generally be explained by means of local computational processes. Since abductive inferences (and also learning and perception[35]) are global, non-demonstrative, non-modular processes,[36] Fodor is skeptical of CTM's ability to explain them.

If Fodor is right and CTM can only explain modular and local (usually demonstrative) mental processes, its scope ends up being rather limited. If CTM does not account for global (or nondemonstrative) processes, it ends up leaving unexplained some of the mental processes that are "most characteristic of human cognition." (FODOR, 2001b, p. 05). As far as reasoning is concerned, we don't always, and perhaps don't even usually, think by following demonstrative rules of inference.[37] Even when our mental processes are rational, as opposed to being arbitrary associations of content, they often seem to have the form of inductive inferences, where the conclusion doesn't follow necessarily from the premises. Fodor is certainly right in saying that

---

[35] Fodor clearly takes perception to involve non-demonstrative inferential processes, but his position as to whether it is modular is not so clear. In *The modularity of the mind* he explicitly says that perception is not encapsulated: "Nobody doubts that the information that input systems provide must somehow be reconciled with the subject's background knowledge. We sometimes know that the world can't really be the way that it looks, and such cases may legitimately be described as the correction of input analyses by top-down information flow. (This, ultimately, is the reason for refusing to identify input analysis with perception. The point of perception is the fixation of belief, and the fixation of belief is a *conservative* process – one that is sensitive, in a variety of ways, to what the perceiver already knows. Input analysis may be informationally encapsulated, but perception surely is not.)" (FODOR, 1983, p. 73). However, in "Observation reconsidered", he argues that perception can be inferential, with access to background information, while still being "encapsulated from much of what the perceiver believes." (FODOR, 1990, p. 244).

[36] Demonstrative inferences are treatable by CTM, since they take into account only local properties of mental representations, but I do not think they should thereby be considered modular. This is because valid inferences can operate on representations from any domain. As Bermúdez notes, "what is distinctive about this sort of inference [*modus tollens*] is that it makes no difference what sentences one puts in place of *A* and *B*. In the standard terminology, this inferential transition is *domain-general*." (2014, p. 99). As far as I know, Fodor says nothing about this. I think, though, that demonstrative inferences are a special case of central processes, which are domain general but also local, because they involve transformations which are only sensitive to intrinsic properties of representations.

[37] There are famous experiments in the psychological study of reasoning (e.g. Wason selection task) which can be interpreted as showing that our thought processes are not as similar to valid arguments as Fodor at times seems to think.

rational processes are a fundamental part of cognition, and I also think he is right to say that our thought processes frequently resemble arguments. But these arguments are usually non-demonstrative. We may often come to true conclusions from true premises, but this passage seems to occur typically through inductive inferences, and not deductive ones. If we accept that non-demonstrative inferences cannot be treated as transformations of symbols sensitive only to their local syntactic forms, we must conclude that not even the class of rational mental processes is fully explained by the computational theory of mind. In fact, not even the thought processes typical of Sherlock Holmes, which is Fodor's favorite example, can be used as examples of mental processes vindicated by CTM. This is because it is questionable that even Sherlock Holmes thinks mainly deductively. Although Sherlock Holmes himself calls his thought processes deductive, in many cases the sort of inference he makes would be best described as inferences to the best explanation.

In short, offering psychologically plausible computational explanations only to a limited number of mental phenomena can be taken as a limitation of CTM. As Fodor conceives it, CTM ends up showing only how local mental processes are mechanically possible, but not global mental processes. If Fodor is right and abductive inferences are global processes, mechanistic explanations that follow CTM could not be given for most rational mental processes. So CTM ends up having an even smaller scope, not accounting for associations or for non-demonstrative rational processes.

But even if it is true that only demonstrative inferences and modular processes can be implemented by computational processes, this is in a sense no small matter, since it is usually assumed that several aspects of the mind are modular. In fact, some have argued, against Fodor, that there are no global, or central processes. The mind is, as it is sometimes put, massively modular. If this is right, then all mental processes can somehow be implemented by local computational processes. Peter Carruthers (2006), for example, says that abductive inferences could be accounted for assuming that the mind is massively modular, as long as we assume that modules can interact with each other. Another way out of Fodor's objections to a computational approach to the mind is to adopt a different notion of computation to explain the phenomena that are not explained by the classical model. Pinker (1997, 2005), for example, accepts both massive modularity and the idea that classical computation does not offer a full explanation of mind. He thinks Fodor is taking Turing machine computationalism too seriously, and that other sorts of computations, like the ones accepted by hybrid models of cognitive architecture (which

use both symbolic representations and connectionist representations) can be brought to bear on the attempt to explain abductive processes. So it is not completely unreasonable to be skeptical about Fodor's skepticism about CTM.

3.3 Some other critical remarks

We have seen that behind the ambition to explain mental processes in purely syntactic terms is the intention to understand how rational mental processes can occur mechanically. To suppose that mental processes are computational processes is a good way of explaining how mental processes typically preserve semantic relations such as truth, while still being material processes. As Fodor notes, syntactic properties can show us the path to the physical, mechanical implementation of semantic properties of symbols, to the extent that there is a parallelism between syntax and semantics, that is, that semantic properties can be preserved by syntactic properties. In Fodor's view, computational processes are a first step in the implementation of mental processes involving intentional states, which makes them less mysterious:

> we know that computations can be implemented by causal mechanisms; that's what computers do for a living. So, the thesis that mental processes are computations was not committed to the dualist metaphysics for which earlier versions of mentalism had frequently been chastised. (FODOR, 2008, pp. 104-5).

But there are at least two other critical remarks that can be made against this approach. First, it is not clear that computational processes can give us an exhaustive explanation of demonstrative mental processes. Let us accept that we sometimes think by following the rule of modus ponens, for example. So at least sometimes, when I think that *A* and that *If A then B*, I end this sequence by thinking *B*. Still, saying that mental processes are computational processes does not explain why we tend to apply the rule of modus ponens to representations *A* and *B* that are coherently related. In theory, we can fill *A* and *B* with anything. But in fact we tend not to have sequences of thoughts like "Triangles have three sides", "If triangles have three sides, then penguins do not live in Brazil", "Therefore, penguins do not live in Brazil." When we think that *If*

A then B, in general *B* represents something that is somehow related to what *A* represents (a consequence of it, for example). Logic is one thing, psychology is another. So if what we want to explain are actual mental processes, something more will have to be said about the semantic coherence of mental processes. Fodor gives us no indication of how a computational theory might explain this phenomenon. CTM assumes that the only thing that plays a causal role in mental processes is the syntactic form, not the content of mental representations. But even when we have a sequence of thoughts that may be in part so explained, it is unlikely that this can be its *complete* explanation, inasmuch as it lacks an explanation as to why these transformations involve representations which maintain other relations among themselves, like being representations of things that are cause and effect, etc. There may be some syntactic correlate that preserves these other relations among the contents of mental representations, but we have seen that Fodor himself is skeptical of CTM's ability to deal with non-local, context-sensitive relations. So it is curious that he doesn't note this limitation of CTM, even when it comes to explaining demonstrative processes.

There is also a second criticism that can be made, this time against the assumption that without computational processes, we are left with no prospect of offering a naturalistic account of intentional mental processes. If Fodor is right and there is no plausible parallelism between syntax and semantics in global mental processes, the syntactic level could not provide the path to the semantic implementation of these processes. But that doesn't mean we would be thereby committed to substance dualism, having to assume that mental states don't belong to the natural order. The transition from one mental state to another will not be completely mysterious if there is no computational explanation of them, because we can always go another level down, and assume that there will eventually be a naturalistic explanation captured at the neurological level. So the falsity of computationalism does not imply the truth of substance dualism. Given that mental processes occur in the brain, it is only natural to assume that we will eventually have at least neuroscientific explanations of how they occur. Neural explanations could give us some minimal insight on how states with semantic content can have causal power – they require a structure like the brain to do it. So if we want to know how coherent reasoning is mechanically possible, we can investigate what happens in the brain when we perform these rational mental processes, since the brain is in fact our best example of a device capable of processing information, coherently or not. I'm not saying this is an easy task, I'm just saying that, even if it

ends up turning out that cognitive psychology makes ineliminable reference to intentional states causing others qua intentional states (with no purely syntactic transformations involved), we can at least hope that neuroscientists will decipher whatever happens in the brain when these psychological processes occur. I don't see why this wouldn't be enough to accept that mental states are part of the natural realm.

In fact, what I'm saying is compatible with Fodor's (1975) idea that, for the processes to which CTM offers no explanation, we can still expect that cognitive psychologists will at least *describe* them in intentional terms. Given that we expect them to be *explained* by mechanisms, we can leave their explanation to a lower level science, such as neuroscience. That is, if not all that goes on in the mind can be treated computationally, cognitive psychology could still be responsible for investigating and formulating intentional laws, in which mental states cause others by virtue of their semantic contents. So even if not all psychological laws are implementable by computational processes, we need not suppose that they somehow violate naturalism, as long as we accept that intentional states and mental processes are performed in the brain, and that neuroscience will eventually reveal to us the mechanisms of their implementation.

Perhaps a more appropriate way to delineate the scope of application of cognitive psychology would be to say that it investigates all that goes on in the mind at a level below common sense psychology, and at a level above neuroscience. This may or may not involve the specification of computational mechanisms involving local syntactic transformation of symbols. Fodor would probably say, though, that if cognitive psychology can't provide computational processes that implement psychological laws, it is only describing, rather than explaining intentional mental processes, since he assumes that in order to explain mental processes, we need a mechanism. But still, my point is that Fodor does not need to assume, as he seems to in later work, that coherent mental processes would be mysterious or contrary to a naturalistic view of the mind if they were not explained in terms of symbol manipulation, since he himself acknowledges in *The language of thought* that neural explanations could still be given.

We can, then, legitimately ask whether we really need a symbolic and computational level of explanation of cognition. Considering that the scope of CTM, as Fodor conceives it, is limited, would it not be simpler to offer only intentional descriptions and neuroscientific mechanistic explanations of all mental processes, without the intermediation of computational explanations? I think not, for the assumption that mental processes are computational is a foundational

assumption for the work being done in the cognitive sciences, and it is, at a minimum, useful for the explanation of modular mental processes and demonstrative reasoning. Simply specifying what goes on at the neural level, without looking at the information processing that is going on, though perhaps sufficient for the purpose of offering a naturalistic account of the mind, would be in a sense an uninformative endeavor (just like explaining the behavior of a computer by its hardware only, with no reference to the programs it runs). It would be difficult to see how what happens in the brain relates to what happens in the mind. Some even consider the neural level and the computational level of explanation of the mind to be deeply intertwined. Boone and Piccinini (2016), for instance, argue that the computational level is not independent from its mechanisms of implementation. What cognitive neuroscientists assume, according to them, is that the brain performs neural computations over neural representations. So the working of the brain is itself understood in light of information processing, and not simply in biological terms.[38]

In any case, I'm more interested in considering the language of thought hypothesis, and not so much its role in a computational theory of mind. So even if CTM proves to have the limited range of application that Fodor thinks it has, or even if we accept that not all mental computations are over symbols in a language of thought, there are other good reasons for accepting that there is a language of thought. As we've seen, it seems reasonable to assume that propositional attitudes are relations with mental representations. We will see in chapter 3 the arguments from productivity and systematicity, which suggest that these representations are structured, like sentences in a natural language. The language of thought, insofar as it consists of syntactically and semantically structured mental representations that are different from any natural language, captures a lower level of explanation than the level captured by common-sense psychology. It is not a simple description of how states with semantic content relate to one another, but rather an explanation of some of the essential characteristics of propositional attitudes. Thus, the arguments we will see in chapter 3 help to support the idea that we need a symbolic level of explanation of the mind even if CTM does not prove to be the complete truth about mental processes. In the following chapter we will contrast RTM and CTM with some opposing views of the mind.

[38] The issue of the autonomy of psychological explanations will be briefly addressed in chapter 2.

CHAPTER 2

OPPOSING THEORIES AND ADVANTAGES OF RTM AND CTM

> "nothing ever reduces to anything, however hard philosophers may try"
>
> (Fodor, 1998c, p. 66)

We saw in the first chapter the central aspects of the representational and computational theories of mind, and some of their limitations. A question that arises, then, is what their rivals are, and what the advantages are of accepting them. That's the topic of this chapter. In part 1, I will show how RTM is opposed to philosophical behaviorism. In part 2, I discuss some problems of the identity theory, which holds that mental states are identical with brain states. In part 3, I consider some of Searle's remarks against CTM. This exposition is not intended to be an exhaustive treatment of behaviorism, the identity theory or Searle's theory. My intention is merely to indicate how RTM and CTM differ from other approaches to the mind, and to show that they seem to have some advantages over the opposing views.

Let us first look at some central aspects of the representational theory of mind, and then briefly at what some competing theories say, and why they do not seem so advantageous. At first, being a realist theory about the existence of propositional attitudes, RTM contrasts with eliminativism, which denies the existence of intentional states. The paradigmatic example here is the eliminative materialism of the Churchlands (which I will not discuss here).[39] In addition to being a realist theory, RTM can be classified as a materialist theory of the mind. In principle, as Fodor (1981a) admits, functionalism – and possibly also RTM and CTM – is compatible with substance dualism (the Cartesian idea that mind and matter are two different substances) and even with idealism (such as Berkeley's, according to which there is no matter, only minds and ideas). But to the extent that we are interested in a naturalist theory of the mind, there is no need to assume any of these views, for RTM and CTM are also compatible with a materialist view of the mind. In addition, functionalist theories are typically non-reductive about mental states. As

[39] Galen Strawson describes, in my view aptly, the adherents of eliminativism by saying that they, by being "passionately committed to the idea that everything is physical, make the most extraordinary move that has ever been made in the history of human thought. They deny the existence of consciousness." (STRAWSON, 2016).

we have seen in chapter 1, a theory that vindicates intentional common-sense psychology must accept entities that are essentially semantically evaluable and causally efficacious. In taking propositional attitudes to be functional relations that we have with mental representations, and in taking mental processes to be computations, RTM and CTM preserve both the semantic aspect and the causal aspect of propositional attitudes. Hence, they do not reduce mental states and processes, as they are accepted by common sense, to entities without these properties.

As a non-reductive materialist theory of mind, Fodor (1975, 1981a) notes[40] that RTM opposes materialist theories that can be considered reductionist, such as behaviorism and the identity theory (also known as type-physicalism). Behaviorism can be characterized as the theory according to which mental states can be reduced to behavioral dispositions. The identity theory can be characterized as the view that types of mental states are identical to types of states in the brain. As Fodor notes, these theories can be considered forms of materialism, since mental states are nothing more than material entities: behavior (or behavioral dispositions), in the case of behaviorism, brain states in the case of the identity theory. They do not, unlike dualism, take mental states to belong to a domain other than the domain of material entities. But unlike RTM, they can be considered reductionist, since for them mental states do not have an independent status, but are instead reduced to other types of entities: behaviors and brain states.[41] That is, both behaviorism and the identity theory seem to reduce common-sense beliefs and desires to entities that do not preserve the two characteristics that, according to Fodor, are essential to them: causal efficacy and the possibility of being semantically evaluated. Thus, an initial problem for both behaviorism and the identity theory is that, insofar as they are reductionist theories, they can hardly be said to vindicate common sense intentional psychology.

Let us see why behaviors and brain states do not seem to preserve these two characteristics. As for causal efficacy, it could in theory be preserved by brain states, as well as by stimuli and behavioral dispositions, insofar as these are material entities. An identity theorist could explain the causation among mental states by saying that it is nothing more than a brain state causing another brain state. A mental state causing a behavior would similarly be nothing more than a brain state causing a behavior. So the identity theory passes this first condition for

---

[40] The exposition in this and the next paragraph is inspired by Fodor (1975, 1981a).

[41] Perhaps both behaviorism and the identity theory can be considered realist theories of mental states, in the sense of not being eliminativists (of not simply denying the existence of mental states). According to these theories, mental states exist, they just are not exactly what we think they are; they are reduced to other types of entities: dispositions of behavior, in the behaviorist's case, and brain states, in the case of the identity theorist.

the acceptance of propositional attitudes, since neural states are legitimate candidates to be causes of other states and behaviors. Behaviorism might initially try to account for some cases of mental causation by saying that a mental state causing a behavior is nothing more than the disposition to produce a response given a certain stimulus. For example, instead of saying that it was thirst that caused John to drink water, Fodor (1981a) notes that the behaviorist might say that being thirsty is the same as being disposed to drink water if water were available, and that in this case there was water available. There was a stimulus, the presence of water, which led John to a respond by drinking the water. However, it is questionable that this behaviorist approach actually preserves the common sense notion that mental states cause behavior. As we shall see, Fodor argues that this behaviorist approach to causation from mental state to behavior, only in terms of dispositions, ignores that we often consider that mental states, by virtue of their interactions with other mental states, literally *cause* behaviors; that behaviors are not mere responses to stimuli, but are intermediated by mental states. Additionally, Fodor notes, for example, that behaviorism does not seem to satisfactorily explain causation among mental states (or mental processes), such as reasoning, in purely behavioral terms. Thus, stimuli and behavioral responses cannot be regarded as causally efficacious in the same sense as mental states, since they do not account for causation among mental states, and doubtfully preserve the idea that mental states cause behavior. The entities accepted by behaviorism do not seem to be sufficient to account for all kinds of mental causation admitted by common sense.

As for the second characteristic common sense attributes to mental states, presumably behaviors and brain states are not semantically evaluable. It does not seem appropriate to say that brain states or behaviors can be true or false (like beliefs), or fulfilled or frustrated (like desires). It is at a minimum questionable that semantic content can be attributed to brain states and behavioral dispositions, just as we can say that a belief or a desire is about something, or is directed to something external to it.[42] And if so, these theories of mind probably cannot be taken to be entirely realist theories about the states presupposed by intentional common-sense psychology, since they reduce such states to entities which do not preserve all their essential properties. In this respect, they differ from Fodor's representational theory. Since they do not accept entities with the same essential properties as the ones propositional attitudes have for common sense (i.e. causal efficacy and semantic evaluability), it does not seem possible to say that

[42] I do admit that the case is less obvious for neural states, since it is usual now, in cognitive neuroscience, to talk about neural representations (see Boone and Piccinini, 2016).

they vindicate common-sense psychology, as RTM does. If we consider it advantageous for a theory to preserve the typical characteristics of mental states, which are part of our pervasive practice of explaining and predicting behaviors, then RTM has this advantage over behaviorism and the identity theory.

Let us see in a little more detail some of the problems of these two theories, first behaviorism, then the identity theory. In the third part of this chapter, I will introduce some of Searle's criticisms of computational theories of the mind, and show how one could respond to them.

1. Behaviorism

"ontology is one thing, epistemology is quite another."
(Fodor, 2003c)

Hilary Putnam (1963; 1967) was perhaps the first philosopher to oppose both behaviorism and the identity theory. He raised serious problems for the behaviorist hypothesis that mental states, like pain, are behavior dispositions. Putnam acknowledges that the behaviorist's conception of pain has the advantage of being more in line with how we verify that someone is in pain. When we attribute pain to someone, we usually do so based on the person's behavior, not by observing her functional organization, or her brain. However, while the way we verify whether someone is in pain may tell us something relevant about the concept "pain", Putnam rightly observes that it does not tell us what pain is. In general, the way we verify that a thing x is A may in fact be irrelevant for knowing what property A is. To use one of Fodor's expressions, to think that verification criteria determine what a thing is would be to "put the epistemological cart before the ontological horse."[43]

In everyday life, I don't verify that a substance is water by analyzing its chemical composition, but that does not mean that water is not H_2O. Similarly, I certainly do not check that someone is in pain by opening her skull, or by analyzing her functional organization, but that

[43] Fodor says this in the context of an objection to Kim (1997, p. 13).

does not mean that pain is not a cerebral or a functional state.[44] In "Brains and behavior" (1963), a paper in which Putnam, in his own words, attempts to bury logical behaviorism, he offers more reasons for not wanting to identify pain with a disposition to behave in a certain way. He argues that from the fact that we can conceive of worlds in which pain does not correspond to any kind of behavioral disposition we can conclude that there is no necessary connection between pain and typical pain behavior, and therefore that pains (and presumably other mental states) cannot be reduced to behavior dispositions. Moreover, according to Putnam, the word "pain" does not mean a certain group of behavioral responses to certain stimuli, but rather "the presence of an event or condition that normally causes these responses." Thus, against logical behaviorism, he concludes that sentences about pain cannot be translated into sentences about behaviors without loss of meaning, because by "pain" we mean what *causes* certain typical behaviors, not the behaviors themselves.

Fodor also notes that we not only take behaviors to be effects of mental states, but also that behaviors are often the result of mental processes, that is, of causal sequences of mental states. So it doesn't seem possible to associate a single mental state with a behavior disposition, because it is common for the *interaction* of different mental states to give rise to a behavior. According to him,

> Mental causes typically give rise to behavioral effects by virtue of their interaction with other mental causes. For example, having a headache causes a disposition to take aspirin only if one also has the desire to get rid of the headache, the belief that aspirin exists, the belief that taking aspirin reduces headaches and so on. Since mental states interact in generating behavior, it will be necessary to find a construal of psychological explanations that posits mental processes: causal sequences of mental events. It is this construal that logical behaviorism fails to provide. (FODOR, 1981a, p. 116).

The behaviorist does not seem to be able to explain, in purely dispositional terms and without mentioning mental states, how one can, for instance, have pain and not manifest it *because* one believes that it is shameful to express pain. Even if it were appropriate to talk about

[44] Place (1956) presents a similar argument to say that there is nothing conceptually problematic about the hypothesis that consciousness is a brain process. He makes it clear that, with this hypothesis, one does not want to give a definition of what "consciousness" means. That is, he is not saying that sentences about sensations are reducible or analyzable in terms of sentences about brain states. The "is" in the sentence "consciousness is a brain process" is what he calls an "is" of composition, analogous to the "is" in "lightning is a motion of electric charges."

behavioral dispositions, these dispositions themselves are the result of the interaction of mental states, and not something to which we can reduce mental states. If we consider propositional attitudes, reducing them to behavior dispositions seems even more implausible than reducing sensations like pain and thirst. I think the same objection to the functionalist characterization of the desire to eat chocolate applies here. After all, to what kind of behavioral disposition can we reduce the belief that Venus is closer to the sun than the Earth, or the desire to eat chocolate? If someone asks Mary whether it is Venus or Earth that is closer to the sun, she may decline answering because, e.g., she doesn't like the sound of her own voice. But that doesn't mean she doesn't have that belief. Having the desire to eat chocolate will only be associated with being disposed to eat chocolate if one doesn't simultaneously have the desire not to harm one's teeth, or the desire to lose weight coupled with the determination to resist the temptation of chocolate. That is, it does not seem possible to characterize the belief that Venus is closer to the sun than the Earth, or the desire to eat chocolate, in terms of dispositions to behave in a certain way without other mental states infiltrating in the characterization of the disposition, and therefore without eliminating mental states from explanations of behavior.

Even if the behaviorist were successful in reducing mental states to behavioral dispositions, leaving the mental state out is leaving the cause of the behavior out, as Putnam observes. That is, even if there were a correspondence between mental states and behavioral dispositions, which there doesn't seem to be, it is not clear that we should be satisfied with this characterization of mental states, since in so charactering them we lose the explanatory power we ordinarily attribute to mental states, but not to behaviors. Fodor observes that behaviorism invites us to deny the undeniable: the contribution of internal states to the causation of behavior (cf. FODOR, 1975, pp. 01-02). It is unclear how we could predict and explain behavior with no reference to mental states.

One example of a philosophical approach that, in the spirit of behaviorism, takes mental states and processes to be explanatorily irrelevant, at least when it comes to explaining linguistic behavior, is found in Wittgenstein. In *The blue book*, for instance, Wittgenstein criticizes the idea that it is necessary to assume the existence of mental processes that give life or meaning to linguistic signs. He especially criticizes the idea that we must form mental images that interpret linguistic signs, in order to understand a sentence. For example, if I ask someone to bring me a

red flower, according to Wittgenstein, it is not necessary to suppose that a mental image of a red flower passes through his mind, serving as an interpretation of the request and allowing him to execute it. Wittgenstein's point, I think, is that a mental event that functions as an interpretation of linguistic signs does not always occur, and need not occur, for words to be meaningful, or for us to be able to, e.g., follow orders. Even if a mental image of a red flower does occur when I hear the order to pick up a red flower, for Wittgenstein this image cannot be what explains my understanding of the order. Were we to suppose that what explains the comprehension of a linguistic sign is a mental sign, we would then be required to say what in turn explains the understanding of that mental sign. If it were another mental sign, the problem would go on ad infinitum. Wittgenstein then suggests that the meaning of a word is not something mental accompanying it, but rather is its use. As he says,

> The signs of our language seem dead without these mental processes [of understanding and meaning]; and it might seem that the only function of the signs is to induce such processes, and that these are the things we ought really to be interested in. (...) But if we had to name anything which is the life of the sign, we should have to say that it was its *use*. (WITTGENSTEIN, 1958, p. 03-04).

For Wittgenstein, the linguistic sign is not something dead that gains meaning through a psychological process of understanding, which takes place in a mysterious mental medium; it gains meaning or life by being used in a certain way. Mental processes are, for Wittgenstein, unnecessary to explain behaviors such as the obedience to an order.

I think it is possible to raise against Wittgenstein similar problems to those that we have raised against the behaviorist characterization of pain. The meanings of the words in a language can be established by their use, but it is questionable that understanding their meaning *is* the same thing as knowing how to use them. We can apply to Wittgenstein's flower example the distinction drawn by Putnam between what a thing is, and the ways or criteria we use to verify it. There is no need to deny that we would verify that someone has understood an order to bring a red flower by observing his or her flower-picking behavior, but that does not mean that the understanding of that order is the same thing as being disposed to bring a flower. Similarly, we typically know if someone understands a sentence by observing if the person uses it correctly. But that doesn't mean that using a sentence correctly is all there is to understanding it. On the

contrary, using a sentence properly, or obeying an order, as Putnam might say, seems to be a *consequence* of the understanding of a sentence or an order.

If nothing needs to go through the head of someone who is asked to bring a red flower, how can we explain, for example, the differences in behavior between a Portuguese speaker and a monolingual Chinese speaker when receiving a request in Portuguese? It seems reasonable to say that the Portuguese speaker brings the red flower in part *because* he understood the order, while the Chinese speaker remained static *because* he did not understand the order. To the extent that the understanding of an order is part of the causal chain that leads one to bringing a red flower when asked to, it seems false that no underlying mental processes are required to explain behavior.

It is true that I don't verify that one feels pain, understands a word, etc., by observing what's going on in their mind; I do not check a person's mental state before attributing her pain or understanding. But this does not mean that the understanding of a word, pain, etc., *are* behavioral rather than mental events. When I say that someone is in pain, I do not mean to be saying that she is behaving in a certain way. What I mean is that she is *feeling* pain, and that this feeling is the cause of her pain behavior, even if the behavior (and not the feeling) is the evidence I use to attribute pain to her. Likewise, that someone uses the words of a language correctly, and that she brings me a red flower when requested, may be the evidence I use to determine that this person understands what I say. But understanding a request is not the same thing as being disposed to obey it.

Understanding an order cannot be the same thing as being disposed to obey it because there is no necessary connection between one and the other, just as there is no necessary connection between feeling pain and being disposed to exhibit pain behavior. One can understand an order and not be willing to obey it, or obey it by chance, without understanding it. It is possible to imagine a situation, similar to Searle's Chinese room (which we will see in more detail in section 3 of this chapter), in which someone who does not speak Portuguese systematically brings me a red flower whenever I ask for one in Portuguese, because of an incredible coincidence, without her action being caused by my request, and therefore without her having *understood* my request. If being disposed to obey the order were the same thing as understanding it, we should say in this case that the person understood the order, which seems absurd. If this is so, it seems reasonable to say that there is a difference between behavior and

linguistic understanding, the former being generally a consequence of the latter.[45] And if linguistic understanding is, as it seems to be, a mental event, then mental processes corresponding to the interpretation of linguistic signs are not unnecessary or irrelevant to explain behavior, as Wittgenstein seems to believe. Mental events are typically causes of behavior, including language related behavior.

Wittgenstein sometimes seems to think that, if mental processes were involved in the understanding of a sentence, they would take the form of mental *images*. He then opposes this idea. His main point seems to be that these images are not necessary to explain linguistic understanding.[46] But with that I'm in perfect agreement. In saying that linguistic understanding involves mental processes, and that it is not just a matter of how we act given certain linguistic stimuli, I don't mean to be saying that understanding a sentence requires forming conscious mental *images*. Most cognitive scientists, including Fodor, would deny that, along with Wittgenstein. But it would be wrong to conclude, from the fact that there need be no conscious mental images accompanying linguistic understanding, that no mental processes are involved in linguistic understanding. What is being assumed here is that an explanation of the understanding of a sentence will appeal to mental states and processes, and the science of Psycholinguistics would not have come so far without this assumption. But it is not being assumed that these processes need to have images as their vehicle.

Of course Wittgenstein might legitimately and pertinently demand an explanation of this mental understanding that gives life to linguistic sings, especially if it occurs in the form of another sign, this time a mental sign, which in turn needs to be interpreted. That is, if the signs of a language acquire life (or meaning) because they express mental signs, we can ask what gives life to these mental signs. Do we need postulate other signs that give life to them, leading to an infinite regress? I take that this is one of his main points against mentalist accounts of linguistic understanding. This may in fact be a problem for proponents of the language of thought, who

---

[45] As Fodor remarks, "if *anything* is clear it is that understanding a word (predicate, sentence, language) isn't a matter of how one behaves or how one is disposed to behave. Behavior, and behavioral disposition, are determined by the interactions of a variety of psychological variables (what one believes, what one wants, what one remembers, what one is attending to, etc.). Hence, in general, any behavior whatever is compatible with understanding, or failing to understand, any predicate whatever. Pay me enough and I will stand on my head iff you say 'chair'. But I know what 'is a chair' means all the same." (FODOR, 1975, p. 63).

[46] It is possible that Wittgenstein would argue against the view that linguistic understanding involves mental images based on introspection, claiming that we don't usually visualize a red flower when we hear and understand the request to bring a red flower. But using introspection to deny that there are any mental processes involved in understanding a sentence is a hasty move, since cognitive science works under the assumption that there are lots of unconscious mental processes that explain behavior.

accept that there are mental (though not necessarily imagistic) symbols. But, as Cain notes, "it is far from clear that cognitive scientific explanations must be either circular or lead to infinite regress." (CAIN, 2002, p. 40). He argues that the regress can be eliminated if we accept that CTM shows us how the manipulation of symbols in the mind can happen in a purely mechanical way. Natural language symbols can derive their meanings from mental symbols, which in turn are assumed to be physically realized in the brain. Mental symbols would end the symbolic chain, with no need to assume more symbols to explain our understanding of them. One could then simply say that the symbols in the language of thought don't require any interpretation. We interpret natural language sentences, but we simply have thoughts, we don't interpret our thoughts.[47]

I'm not trying to say that the semantics of the language of thought is not a problematic topic. Here one could still demand some explanation for the semantics of mental symbols. Fodor does attempt to give one, as we will briefly see in chapter 4. But from the fact that we need to account for the semantics of mental symbols it doesn't follow that these symbols do not exist, nor that assuming their existence is not explanatorily fruitful.[48] Language related behaviors would be mysterious if we did not attributed mental causes to them. There does not seem to be anything absurd about the assumption that mental events cause behavior, since we daily assign mental causes to explain and predict other people's behavior.

As distant as Wittgenstein's and Fodor's ideas may be, there being perhaps nothing (the second) Wittgenstein would more strongly object to than the idea of a language of thought, it is interesting to note that their philosophical projects seem to have similar motivations. Wittgenstein makes it clear that he wants to reject the notion of mind as an immaterial entity whose mechanisms are incomprehensible, as well as the notion of thought as a mysterious attribute belonging to this "queer kind of medium" (1958, p. 03). Ryle (1949), in a way, has the same concern. He criticizes Cartesian dualism, which would have the undesirable consequence that we cannot know the mental states of other people, since the body is different from the mind and the mind is not directly accessible. The solution Ryle found to preserve the intuitive idea that we can know the mental states of others was roughly to say that there is nothing beyond what can be

[47] Fodor does in fact give a similar answer when considering an objection against one of his argument for LOT: "My view is that you can't learn a language unless you already *know* one. It isn't that you can't learn a language unless you already *learned* one." (FODOR, 1975, p. 65).

[48] I will come back to this topic in chapter 4, where I present the arguments for assuming that thought is independent of language, and that language is used to express thought.

shown in behavior. What bothered Ryle and Wittgenstein, and what they wanted to deny, was primarily the idea that mental states are private entities, only knowable to the person who has them.[49]

But from what we have seen so far, Fodor would be in perfect agreement with the rejection of Cartesian dualism. When Fodor compares the mind to a computer, what he wants is precisely to bring it back to earth. Mental states, like anything else, must be explainable from a naturalistic point of view, and therefore without mysteries. The difference is that Wittgenstein, in attempting to bring the mind back to earth, puts it, so to speak, in behavior, stating, for example, that when we write, we think with our hands, or that when we speak, we think with our mouths and the larynx (see WITTGENSTEIN, 1958, p. 06). Fodor takes thoughts back to their natural place, namely, the mind, and proposes to remove the air of mystery that surrounds intentional mental states by treating them as involving symbols physically instantiated in the brain. Fodor observes (1975, p. 04) that behaviorists like Ryle (and presumably Wittgenstein, as I read him) assume that the acceptance of mental states implies the acceptance of a dualism, and therefore of a conception of the mind as something private and mysterious. The only acceptable alternative would be behaviorism. But this is indeed a false dilemma, since RTM can be conceived as a mentalist but non-dualistic theory (about substance). There is no reason to suppose that we have no way of knowing other people's mental states – a skeptic would certainly pressure Fodor at this point, but there is no reason to suppose that philosophy of mind should be guided by epistemological precepts (as was commonly assumed).

So I agree with Fodor that RTM is a better theory, for it tries to preserve our common conceptions of beliefs and desires as states with semantic content and causal power. Wittgenstein's strategy, insofar as it avoids reference to mental processes, is less convincing for the reasons given above. As Fodor says, "there are various things that you can usefully do when your car gets a ping in its cylinders; but declining to quantify over the engine is not among them. You need a story about the engine to explain how the car behaves." (1987, p. xi). If you car's engine stops working,

[49] I do not intend to explore here Wittgenstein's private language argument. Suffice to say that Wittgenstein's concerns, as with most philosophers of the early twentieth century, are mainly epistemological. The private language that Wittgenstein finds impossible is a language whose words refer "to what only the speaker can know – to his immediate private sensations. So another person cannot understand the language." (1953, § 243). It is at a minimum questionable that his observations apply to the language of thought, conceived as a construct accepted by cognitive psychologists. Someone who adopts the idea that the vehicle of thought is a language need not also adopt the thesis that its symbols refer to private experiences only knowable to the person who has them.

the best attitude is not to deny or ignore its existence, but to deal with the problem. Likewise, the semantic aspect of mental representations can be especially problematic and require an explanation, but eliminating or ignoring mental representations is not the best answer. With them we eliminate a relevant explanatory level, thus getting an incomplete understanding of reality.

It is important to note that reference to mental states is not present only in our everyday explanations of behavior. The practice of cognitive psychology has proven to be contrary to behaviorist precepts, since mental representations are assumed by most contemporary cognitive theories. As Fodor observes,

> The behaviorist has rejected the appeal to mental representation because it runs counter to his view of the explanatory mechanisms that can figure in psychological theories. Nevertheless, the science of mental representation is now flourishing. The history of science reveals that when a successful theory comes into conflict with a methodological scruple, it is generally the scruple that gives way. Accordingly the functionalist has relaxed the behaviorist constraints on psychological explanations. There is probably no better way to decide what is methodologically permissible in science than by investigating what successful science requires. (FODOR, 1981a, p. 123).

Thus, although behaviorist methodology has played a big role in psychological theories in the past century, it is no longer compatible with the vocabulary of current cognitive psychology and with the kinds of explanation it offers.[50]

As I mentioned earlier, another alternative to RTM is the identity theory. Identity theorists, unlike behaviorists, would not deny that mental states can contribute to the causation of behavior. But given that they accept that mental states are identical or reducible to brain states, they would say that what actually causes my shouting and complaining behavior, for example, is the brain state that constitutes my pain. Although the identity theory succeeds in dealing with this kind of mental causation, it faces other problems. We will look at some of them in the next section.

---

[50] There is currently a movement in philosophy of mind, represented by philosophers such as Daniel Hutto and Erik Myin (2013), who claim that mental representations are unnecessary to explain behavior. They propose what they call a "radical view" about cognition, and assume that it is possible to explain behavior solely by means of the interactions of an organism with its environment. I take this view to be behaviorism in a new guise and, as such, subject to the same criticisms that Fodor and Putnam formulated decades ago.

2. Identity theory, multiple realization and the autonomy of psychology

> "The world, it seems, runs in parallel, at many levels of description. You may find that perplexing; you certainly aren't obliged to like it. But I do think we had all better learn to live with it."
>
> (Fodor, 1997)

In his famous article "Special sciences,"[51] Fodor argues against the view he calls physiological reductionism. Physiological reductionism can be understood as the thesis that the types of states psychology talks about, as well as the explanations and laws that quantify over them, can be identified or reduced to types, explanations, and laws of neurology.[52] If we characterize the identity theory as the thesis that types of psychological states can be identified with types of neurological states (that is, as a form of type physicalism), we can consider that it is presupposed by physiological reductionism, in the way Fodor understands it. An identity theorist would say, for example, that all instances of headache, or of the belief that it is going to rain, are always instances of whatever the cerebral correlates of these states are.[53]

I will consider here that the identity theory is a form of reductionism in regards to psychological types. That is, although it says that there is a relation of identity between psychological and neural types, it seems clear that what is in fact assumed is that psychological types are nothing more than neurological types, thus supposing a primacy of the latter in relation to the former. The identity theory assumes that psychological states do not have an ontological status independent of the ontological status of neural states. If we understand, with Kim, that "to

---

[51] This article reappears, with a few changes, as part of the introduction to the book *The language of thought*. Citations will be from the book version.

[52] In this paper Fodor also argues against a broader reductionist view, according to which all special sciences (basically all sciences other than physics) can ultimately be reduced to physics. The idea would then be that psychology can be reduced to physics, with an intermediate reduction to neuroscience.

[53] Place, an identity theorist, believes that it is possible to analyze cognitive states (such as knowing, believing and understanding) in a behavioristic manner, that is, in terms of behavioral dispositions. However, for him, conscious states, experiences and sensations need to be analyzed with reference to brain processes. So Place (1956) applies the identity theory to conscious states, but not to propositional attitudes. The same goes for Smart (1959), another proponent of the identity theory, who defends the thesis that sensations are brain processes. What Fodor is attacking, on the other hand, is mainly the idea that propositional attitudes can be reduced to brain states. But we shall see that the argument he presents against this idea can also be, to some extent, applied against the reduction of conscious states to neural states.

be reduced is to be eliminated as an *independent* entity." (1992, p. 24), we can characterize the identity theory as a form of reductionism applied to psychological types.

Physiological reductionism can then be characterized as a somewhat broader reductionism, for in addition to adopting the type identity thesis, it also assumes that the ontological status of psychological types has implications for how we interpret the relation between psychology and neuroscience. If psychological types are nothing more than neurological types, it is natural to suppose that it will eventually be possible to reduce the laws of psychology to the laws of neuroscience. That is, if types of mental states such as pains, the belief in P or the desire that Q, are simply identical to types of brain states, nothing prevents us from eventually reducing the explanations of psychology, which quantify over these types, to explanations of brain science. If so, psychology would not be an autonomous science with its own vocabulary, explanations, and laws, since its object of study can be reduced to the object of study of neuroscientists.[54] In "Special sciences," Fodor argues precisely against this idea. According to him, psychology is an autonomous science, which cannot be reduced to lower-level sciences.[55] In what follows, I will present his arguments and critically discuss them.

First, though, it is important to draw a distinction between a physicalism of tokens and a physicalism of types. Fodor's objections are directed against type physicalism, not token physicalism. A token physicalist, in Fodor's formulation, would say that "all the events that the sciences talk about are physical events" (1974, p. 12)[56], so all events that enter into the laws of the so-called special sciences are ultimately events that also have a description in physics, and therefore can be ultimately embraced by the laws of physics. Fodor himself is sympathetic to this idea. He even acknowledges that it is likely that all psychological events currently in existence are

---

[54] As we have seen, in some sense behaviorism also compromises the vocabulary of psychology, as it is conceived nowadays, since its object of study ceases to be intentional or conscious states, and ends up being stimuli and behavioral responses.

[55] In some sense, Fodor's theory can also be considered reductionist, if we conceive it as suggesting that mental states are reduced to functional relations with symbols, and that mental processes are reduced to syntactic driven transformations of symbols. But this kind of information processing, which Fodor sees as being the object of psychology, is closer to the mental states as accepted by common sense, insofar as it preserves their properties of semantic evaluability and causal efficacy.

[56] This thesis appears in a somewhat different formulation in "The mind-body problem": "The identity theory can be held either as a doctrine about mental particulars (John's current pain or Bill's fear of animals) or as a doctrine about mental universals, or properties (having a pain or being afraid of animals). The two doctrines, called respectively token physicalism and type physicalism, differ in strength and plausibility. Token physicalism maintains only that all the mental particulars that happen to exist are neurophysiological, whereas type physicalism makes the more sweeping assertion that all the mental particulars there could possibly be are neurophysiological." (FODOR, 1981a, p. 117).

neurological events (which would ultimately be events covered by the laws of physics). The event consisting of my current belief that it is going to rain may be the same event consisting of a certain network of neurons that is active now in my brain. The tokening of the same belief type in a monkey may be the same event as the tokening of a different network of neurons. In the future, a token of the same belief type may be the same event as the tokening of a certain electrical circuits in a future robot. But one of Fodor's points is precisely that token physicalism does not imply type physicalism, nor reductionism. That is, even if every particular psychological event is also an event that can be embraced by the laws of neuroscience, or of physics, it does not follow that psychological types can be reduced to or identified with neurological or physical types, or that the laws of psychology can be reduced to the laws of neuroscience or physics. Fodor's idea, more generally, is that we can adopt a materialistic stance about the events dealt with by the special sciences, without having to accept that all sciences, including psychology, can be reduced to physics (in the case of psychology, through neuroscience).

Fodor argues against the idea that psychology can be reduced to neuroscience by attacking the identity theory (or type physicalism), according to which types of mental states can be identified with types of brain states. His criticism against the identity theory is inspired by one of the main arguments against type physicalism, initially formulated by Putnam, and known as the multiple realization argument. In "Psychological predicates," Putnam argues that it is highly unlikely that all beings to which we attribute a certain type of psychological state are always tokening the same type of brain state whenever they are in that psychological state. If we accept that pain is identical to a physicochemical state of the brain, then every organism that can feel pain (be it a human being, an octopus or an alien), and only them, must be capable of having the same physicochemical brain state. The same should apply to all kinds of psychological states. Putnam considers this an ambitious and unlikely hypothesis, since there are organisms that have evolved in parallel to us (such as mollusks), and that may not have the same physico-chemical configuration that we have, although they can presumably feel pain. The identity hypothesis will fail, Putnam says, if any kind of psychological state, not just pain, is found to have different physical correlates in different types of organism. Another difficulty for the identity theory is that it does not leave room for the logical possibility that beings who do not have a biological composition, such as robots and possible aliens, may eventually have mental states. Putnam

assumed, then, that it is most likely true that mental states can be realized in multiple types of physical states, and therefore that the identity theory is most likely false.

Fodor basically adopts the same argument to oppose the reduction of psychology to neuroscience, since the identity thesis is presupposed by what he calls physiological reductionism. According to him, psychology would be reduced to neuroscience if all its laws were reduced to laws of neuroscience. Since the laws of a science quantify over the natural kinds of that science, a law of psychology would be reduced to a law of neuroscience if the natural kinds covered by it could be reduced to natural kinds covered by a law of neuroscience. That is, in Fodor's view (1974, p. 11), a necessary and sufficient condition for any law of psychology (formula (1) below) to be reduced to a law of neuroscience (formula (3) below) is that there be bridge laws between the antecedents of psychological laws and the antecedents of laws of neuroscience (formula (2a) below), as well as bridge laws between the consequents of psychological laws and the consequents of neuroscientific laws (formula (2b) below). If all the laws of psychology can be reduced in this way, psychology will have been reduced to neuroscience.

(1) $S1x \rightarrow S2y$
(2a) $S1x \leftrightarrow P1x$
(2b) $S2y \leftrightarrow P2y$
(3) $P1x \rightarrow P2y$

Bridge laws would then be interpreted as saying that psychological types are identical, or coextensive, with neurological types, and this relation of identity, or of coextension, would be a law, that is, it would be true for all possible tokenings of the psychological and neural types in question.[57] Against this, Fodor notes that there is no firm data to prove this coextension. Like Putnam, he finds it likely that the same type of psychological state may occur in different neurological states. He also finds it possible for the same neurological structure to perform different psychological functions at different times, just as computer structures can run different programs at different times. If he is right, then mental types are neither identical nor coextensive

[57] Fodor recognizes that bridge laws can be interpreted as establishing either a relationship of identity or of coextension between natural kinds. According to him, "if reductionism is true, then *every* kind is, or is coextensive with, a physical kind. (Every kind *is* a physical kind if bridge statements express nomologically necessary property identities, and every kind is coextensive with a physical kind if bridge statements express nomologically necessary event identities.)" (FODOR, 1974, p. 14). Identity presupposes coextension, but not vice versa. But this distinction will not be relevant to what follows.

with neural types. But even if they were coextensive, this coextension could not be a law, because, according to Fodor,

> it seems increasingly likely that there are nomologically possible systems other than organisms (viz., automata) which satisfy the kind predicates of psychology but which satisfy no neurological predicates at all. (...) equivalent automata can, in principle, be made out of practically anything. If this observation is correct, then there can be no serious hope that the class of automata whose psychology is effectively identical to that of some organism can be described by *physical* kind predicates. (FODOR, 1974, p. 18).

Thus Fodor's point is that it is unlikely that types of mental states are coextensive with neurological types, but even if they were, such coextension would be only provisional and most likely not a law, if we accept that it is a nomological possibility that there will eventually be robots with psychological states. Neither Fodor nor Putnam refute the identity theory once and for all, insofar as they don't seem to regard multiple realization as a proven phenomenon. But insofar as it is highly probable that mental states are multiply realizable, it is highly unlikely that bridge laws between the types of psychology and the types of neurology are true, and therefore highly unlikely that the identity theory is true. If that is the case, it seems that the predicates and laws of psychology are most likely autonomous with respect to the predicates and laws of neuroscience.

A natural way out for the identity theorist would be to say that all that multiple realization shows, if true, is that mental types are identical not to single neural or physical types, but to *disjunctions* of physical types. The mental type headache, for example, would be identical to a certain neural configuration in humans, another in octopuses, possibly another in aliens, and so on. There would be an identity relation, or reduction, between a kind of psychology and a disjunction of types of neurology or physics. The bridge laws would have on one side one psychological type and on the other side a disjunction of types from more basic sciences. They would then ensure that laws of psychology could be reduced to laws of more basic sciences involving the disjunction of heterogeneous types.

Against this idea, Fodor basically argues that reduction between types can only be achieved if bridge laws have natural kinds at both positions, and the natural kinds of a science are those that are covered by laws of that science. That is, in Fodor's view, the reduction of psychology to a more basic science, like neuroscience, requires bridge laws established between

natural kinds of psychology and natural kinds of neuroscience, each being a kind that is subsumed by the laws of the science to which it belongs. The disjunction of heterogeneous kinds, however, does not seem to be itself a natural kind, because it is not subsumed by the laws of any science. There do not seem to be laws that involve the disjunction (neural state X in human or neural state Y in mollusks or circuits state Z in robots, …) This disjunction could appear at most in some bridge law, but in fact does not appear in laws of a science more basic than psychology. And if that is so, this would not be a legitimate reduction, inasmuch as "*all* that the disjuncts have in common is that they realize some higher level state" (FODOR, 1997, p. 16). Stipulating that a mental type is identical to the disjunction of its realizers would be an arbitrary move, since the disjunction presumably plays no independent role in a lower level science, but only appears in a bridge law whose purpose is to establish that reduction.[58] As Fodor says, these disjunctions are not independently certified, beyond bridge laws (1997, p. 16). In fact, strictly speaking, bridge laws involving disjunctions would not be laws, since Fodor considers that "a necessary condition on a universal generalization being lawlike is that the predicates which constitute its antecedent and consequent should be kind predicates" (1974, p. 20), and disjunctions, he says, are not "kind predicates." Fodor then says that there could be bridge statements involving these disjunctions, but they would not be laws.

In Fodor's view, therefore, psychological natural kinds cannot be reduced to natural kinds that are part of a lower-level science because there is reason to believe that psychological states are multiply realizable, and disjunctions are not good candidates for natural kinds. There can be at most bridge statements of the form $Sx \leftrightarrow (P1x \vee P2x \vee ... Pnx)$, which could be read as "'every event which consists of x's satisfying S is identical with some event which consists of x's satisfying some or other predicate belonging to the disjunction P1 v P2 v … v Pn'" (1974, p. 20). In other words, we could interpret the multiple realization of a psychological type as being compatible with a token physicalism, whereby each event in which x instantiates a mental property S is identical to an event in which x instantiates some physical property.[59] But multiple realization is not compatible with type physicalism, since tokens of the same psychological type are assumed to

[58] The case here is different from the possible situation in which a single psychological type is found to invariably correspond to a *set* (or a conjunction) of neural properties, such that whenever the psychological property is instantiated, a given set of neural properties is instantiated (Aizawa and Gillett, 2011, use this notion of realization). This case is not obviously affected by Fodor's criticism, because the set of neural properties might be an independently certified neural type.

[59] We can understand them as being the same event in the sense that there is only one thing happening, and not two (one event that admits of two descriptions).

be realizable by tokens belonging to heterogeneous physical or neurological types. In Fodor's view, therefore, the multiple realization of psychological states would guarantee that mental properties are autonomous natural kinds, with their own causal powers, not reducible to natural kinds of more basic sciences. There would still be psychological laws about, say, pains, beliefs and desires, even if these states are multiply realizable and there are no laws involving the disjunctions of their physical realizers (since the disjunctions of those realizers would not be part of the laws of any science; they would have nothing in common beyond the fact that they all realize the same psychological type).

Not everyone has accepted this argument, though. A philosopher who questions it is Jaegwon Kim. In the paper "Multiple realization and the metaphysics of reduction" (1992), he argues that multiple realization should be, contrary to what Fodor suggests, interpreted as speaking against the autonomous status of psychology. He thinks that the multiple realization of a mental state should lead us to conclude that the state in question is in fact *not* a single psychological type, about which there can be psychological laws. In order to argue for this, he takes pain as a paradigmatic kind of mental state and compares it with the mineral jade. Kim notes that jade is in fact not a mineral type, since it has been discovered that what we call "jade" is two minerals with a distinct molecular structure: jadeite and nephrite. There do not seem to be laws about jade as such, like "all jade is green," because instances that serve to confirm the law "jadeite is green" do not serve to confirm the law "nephrite is green," and vice versa. If all the instances of jade we had observed so far were in fact instances of jadeite, the law "jadeite is green" would be well confirmed, but not "all jade is green." As Kim says, "jade is a true disjunctive kind, a disjunction of two heterogeneous nomic kinds which, however, is not itself a nomic kind" (1992, p. 12). In other words, there are laws about jadeite, and there are laws about nephrite, but there are no laws about jade as such, because jade is a disjunctive type, which encompasses two different types of minerals. "Jade is green" "is really a disjunction of these two laws: (L_1) "jadeite is green"; (L_2) "nephrite is green"" (KIM, 1992, p. 11).

Kim then asks: why should we not say the same about pain (assuming pain is realized by different physical types)? Why should we not say that, just as jade does not have an autonomous status as a mineral type, and that there are no laws about jade – for jade is nothing more than the disjunction of jadeite and nephrite –, why should we not say that the same is the case for pain? In Kim's view, the multiple realization of pain should lead us to conclude that there is in fact no

natural kind corresponding to "pain," and that there are no psychological laws about pain in general, just as there are no laws about jade. Assuming that pain is realized by the base N_h in humans, N_r in reptiles, and N_m in Martians (and that each of these bases is uniform, i.e. non-disjunctive), Kim argues that there might be theories about pain relative to each species, but no theories about pain in general (because pain would be nothing more than the disjunction of these physical types). For Kim, multiply realizable psychological types are not scientific types because they do not correspond to a single physical type with unique causal powers. If we accept that scientific types are individuated by their causal powers, N_h, N_r and N_m would be heterogeneous types, because they enter into different causal laws. To the extent that the causal powers of an instance of a mental state are identical to the causal powers of its physical basis, a mental state like pain would not have unique and invariable causal powers with respect to its physical basis, for its causal powers are derived from its physical bases (which we are assuming are heterogeneous). Kim comes to the conclusion that "each mental kind is sundered into as many kinds as there are physical realization bases for it, and the psychology as a science with disciplinary unity turns out to be an impossible project" (1992, p. 18). He supports then a local reductionism of psychological types, relative to species. So in Kim's view, multiple realization leads us to the opposite conclusion that Fodor, Putnam, and other antireductionist philosophers intended to derive from it: different physical bases imply different mental types, and to the extent that there are no single psychological types that "survive" the heterogeneous physical bases, psychology would not be an autonomous and irreducible science. Its object of investigation would be in a way dictated by the physical constitution of organisms; different physical bases would constitute different psychological types.

In fact, I think Kim can only draw these consequences from multiple realization because he does not really accept this phenomenon. That is, Kim does not accept that the same type of mental state can be realized by different physical types, which is precisely the idea of multiple realization, because he starts from the assumption that physical types individuate or determine mental types. In "Special sciences: still autonomous after all these years" (1997), published as a response to Kim's paper, Fodor notes this problem and offers a diagnosis of what is wrong with Kim's argument. One of the points he challenges is the analogy between pain and jade. Roughly, Fodor accuses Kim of not drawing a distinction, central to the supporter of multiple realization, between disjunctive properties and properties that are disjunctively realized. Kim assumes that

pain must be like jade (a disjunctive property) because he does not make use of this distinction. For Fodor, jade certainly is an example of a disjunctive property, because it is nothing more than the disjunction of two properties, whereas pain is one property that is disjunctively realized (realized by an open disjunction of different physical bases).[60] Unlike jade, pain is not identical to the disjunction of the properties that realize it. Basically, Fodor thinks Kim does not recognize what's really at stake in the idea of multiple realization. As Fodor says,

> Though Kim *says* that he concedes that psychological properties are MR [multiply realized], that's only because he isn't distinguishing *being MR* (like pain) from *being disjunctive* (like jade). But it's exactly the distinction between *disjunctiveness* and disjunctive *realization* that functionalists are insisting on when they say that pain states are nomologically homogeneous under their functional description *despite the physical heterogeneity of their realizers*. You can't (and Kim can't) refute this claim just by defining "disjunctive kind" so that it isn't true. (FODOR, 1997, p. 13).

In my view, Fodor is right in criticizing the analogy between jade and pain. With this analogy, Kim seems to simply assume that there is no difference between the relation jade maintains with jadeite and nephrite, and the relation pain maintains with its realizers. Kim uses this analogy to say that just as there are no laws about jade (because there are no laws about the disjunction of jadeite and nephrite), likewise there are no laws about pain in general (because there are no laws about the disjunction of its realizers). In doing so, Kim is simply denying the starting point of the proponent of multiple realization, which is that mental states are not identical to the disjunction of their realizers, and that they may be covered by psychological laws as autonomous types, even though they are realized by heterogeneous physical types. If jade is not a property sufficiently similar to pain one cannot infer the inexistence of laws about pain from its multiple realization, as one can infer that there are no laws about jade from the fact that jade is nothing more than the disjunction of two distinct properties. The point of the multiple realization argument is precisely that mental states, *although* multiply realizable, maintain a unity and autonomy. The starting point seems to be precisely that psychology establishes generalizations

[60] In fact, it is important to observe that Fodor follows Kim in using pain as an example of functional and multiply realizable state, although he himself acknowledges that "pain isn't really a happy choice for a working example, since though I take it to be quite a plausible candidate for projectibility, pain is notoriously a very bad candidate for being MR, hence for functional analysis" (1997, p. 10). It is plausible to think that we would deny that a being with a very different physical structure than ours is capable of feeling pain. But this is less plausible in the case of intentional states. Against Fodor, though, I think that a state can have a functional analysis without it being multiply realizable. This point will become clearer later in this section.

involving mental properties whose neural descriptions are somewhat irrelevant, and that the conditions of individuation of these mental properties typically do not require them to be multiplied in order to correspond univocally to however many bases they have. Kim appears to beg the question against the proponent of the multiple realizability argument, for he assumes that it is clear that psychology is dictated by lower level sciences, which is precisely what is at stake (I will come back to this issue in the next subsection).

An advantage, therefore, of accepting RTM and CTM, that is, of accepting that beliefs and desires are functional states involving mental representations, and that mental processes are computational processes, is that they naturally accommodate the possibility of multiple realization; it is left open the possibility that different types of beings, such as octopuses, robots or aliens, may, like us, have mental states like beliefs and desires. They don't take the physical basis of a mental state to be one of its essential properties. Moreover, even restricted to human beings, it seems in fact implausible that there is an identity between mental and neural types, i.e. that headaches, or the belief that it is going to rain, each have exactly the same neural basis in all people, whenever they are instantiated. When we identify mental types with brain types, these possibilities seem to be eliminated, unless we accept that mental types are identical with disjunctions of physical types – a way out that seems unsatisfactory to many because, as we have seen, disjunctions are not natural kinds. Since the computational theory of mind is neutral about the medium in which mental symbols are to be instantiated or realized, it accepts the possibility that mental and neural types are not identical. If for there to be thinking all we need are functionally characterized states which involve representations that can be manipulated computationally (only by virtue of their syntactic forms), and not a particular physical constitution, then the very computations that characterize thinking can in principle occur in beings composed of different materials, or in different neural types. As Fodor says,

> Type physicalism is not a plausible doctrine about mental properties even if token physicalism is right about mental particulars. The problem with type physicalism is that the psychological constitution of a system seems to depend not on its hardware, or physical composition, but on its software, or program. Why should the philosopher dismiss the possibility that silicon-based Martians have pains, assuming that the silicon is properly organized? And why should the philosopher rule out the possibility of machines having beliefs, assuming that the machines are correctly programmed? If it is *logically possible* that Martians and machines could have mental properties, then mental properties and

> neurophysiological processes cannot be identical, however much they may prove to be coextensive." (Fodor, 1981a, p. 117, my emphasis).

The analogy with computer programs is illuminating to understand what is at stake. The same program can be implemented on different types of machines (current computers use silicon, but the same computations could be perform by machines made of valves, or other materials); it is possible to program machines of different materials so that they transform symbols following the same rules. Likewise, Fodor thinks that intentional mental states may presumably exist in different types of materials, provided they realize symbols and a similar functional organization.

It must be made clear, though, that Fodor's theory only says that *intentional* states, such as beliefs and desires, can be characterized in this way (without reference to the physical constitution of the organism or system that has them). We saw in the first chapter that RTM and CTM are theories about propositional attitudes and mental processes involving these attitudes. They say nothing about qualitative states, such as pain. Fodor recognizes that it does not seem possible to give a functionalist account of qualitative states.

We can, if we wish, assume that qualitative mental states can only exist in biological organisms – which seems to me a more defensible position than Putnam's, who thinks that pain is simply a functional state, realizable in any kind of material. But even accepting that sensations, for instance, can only exist in biological organisms, I think it would still be possible to use the multiple realization argument to stop the reduction of psychological explanations involving these states to neural explanations of these states. It would be possible to argue that, even if pain, for example, can only occur in biological organisms, it is still reasonable to suppose that it is multiply realizable in different types of neural states, and, to the extent that it preserves its autonomous status, it would belong to a level of explanation that is not to be confused with the level of explanation of the neurosciences, even if it cannot be functionally characterized.[61] But even if we assume that states with a qualitative character are not multiply realizable, it could still be said that

---

[61] It is possible to draw some distinctions between qualitative and intentional states with respect to their neural bases. As Pereboom and Kornblith note, there is likely to be, across individuals of the same species, a similarity in the physiological structures corresponding to mental states formed by perception. This probability decreases considerably when we consider intentional states. According to them, there is no plausibility to the claim that when you and I believe Baghdad is in Iraq, there is a single physical structure which underlies that belief in each of us. The further we move away from the sensory receptors; the more unlikely we are to find common physical structures underlying our mental states." (1991, p. 718). But to the extent that these sensory or qualitative states correspond to different physiological structures in different species, it is possible to say that they are also multiply realizable.

the generalizations of psychology involving these states talk of pain, for example, as sensations, and not as neural states. I think this would still guarantee the autonomy of psychological explanations in relation to the explanations of neuroscience, as I shall argue in section 2.2.

We have seen then that Fodor assumes that the identity between mental types and neural types is necessary for the reduction of types, explanations, and laws of psychology to types, explanations, and laws of neuroscience. He seems to assume then that in order to stop this reduction one must deny the identity thesis, and that to deny this thesis we need the multiple realization argument. Fodor's argument seems to have the following form:

(1) If psychology is reducible to neuroscience, then types of mental states are identical to types of brain states.

(2) Types of mental states are not identical to types of brain states. [by the argument of multiple realization]

(C) Psychology is not reducible to neuroscience. [by modus tollens]

Unlike Kim, it seems reasonable to believe that multiple realization is sufficient, in the present scenario (but not in every scenario), to deny the identity between mental types and brain types, and therefore to stop the reduction of psychology to neuroscience. And the idea that psychological types are reduced to disjunctions of physical types does not seem satisfactory because, as Fodor says, these disjunctions are not independently certified natural types, which enter into lower-level science laws. But against Fodor, I believe the multiple realization argument is not *necessary* to deny the identity thesis, and stop the reduction, or at least not an epistemological reduction. In fact, we may not even have to deny the identity of mental and neural types in order to preserve some kind of autonomy of psychology. In the following two sub sections I will develop these points.

2.1 On the sufficiency of multiple realization for the autonomy of psychology

Kim seems to believe that multiple realization, in the way he understands it, does not avoid the reduction of mental types to physical types. On the contrary, it speaks in favor of local

reductions, and of dismantling the mental types. One of Kim's points in his article seems to be this: if we accept that pains have different realizers in different species, what authorizes us to say that we are talking about the same pain in all cases? Basically, why should not different physical bases lead us to postulate a multiplicity of mental types, rather than saying that a single type of mental state is realized by different physical bases? Kim believes that physical bases, insofar as they determine the causal powers of mental properties, serve to individuate mental types, which leads him to deny the possibility of multiple realization (understood as preserving the unity of the multiply realizable state). In sum, Kim seems to assume that regularities at higher levels of explanation are always reducible and explained in terms of physical regularities, and that the only entities actually endowed with causal powers are the clearly material entities.

But it could be said that, just as Kim is stipulating that multiple realization should lead us to split psychological types, Fodor is stipulating just the opposite: that psychological types would preserve their unity in the event we find out that they are multiply realized. I think, though, that Fodor offers a more powerful analogy than the one Kim offers with jade. As Fodor says, "the very *existence* of the special sciences testifies to reliable macro level regularities that are realized by mechanisms whose physical substance is quite typically heterogeneous" (1997, p. 21). In other words, the existence of special sciences such as psychology, biology, and geology suggests that there are regularities and natural types at higher levels of explanation that are not captured by regularities involving the natural types of lower level sciences. If there are laws in the special sciences, albeit ceteris paribus, then we have reason to believe that the types involved in these laws have causal power qua types of a particular science (cf. Fodor, 1989). So there is nothing special about psychology in this respect. We shouldn't, then, simply assume that physical (or neural) types determine mental types, just as we don't assume that the individuation of entities of geology, astronomy or economics should be dictated by whatever lower level sciences tell us about their constitution.

In the end, I think it is scientific investigations that will tell us whether mental types maintain a unity despite the heterogeneity of their physical bases, or whether different physical bases should lead us to multiply mental types. Basically, multiple realization is an empirical question. This goes for Kim and Fodor. The best way to know what the status of psychological types is, if we suspect they are multiply realizable, is to observe the practice of psychology and neuroscience, rather than try to determine a priori that heterogeneous physical types can, or

cannot, give rise to homogenous mental types. Would we or would we not revise our psychology, if we were to find different bases for what we take to be the same psychological state? Aizawa and Gillett (2011) try to investigate this. They argue that, in practice, neuroscientists do not always assume that different kinds of neural states give rise to different mental types. According to the authors, the strategy of splitting a mental type when it is discovered that it corresponds to different neural bases *sometimes* occurs (as in the case of the division between declarative and procedural memory, after the discovery that lesions in different areas of the brain can cause the loss of one type or other of memory), but not always.[62]

According to them, there are cases in which psychology and neuroscience consider that the same type of mental state is preserved even after it is discovered that it can be realized by different sets of neural properties.[63] They examine the property of having normal color vision. According to them, it was preserved as a unique higher level property even when studies about the eye recognize that there is variation within the population with normal vision, for example in the composition of green and red photopigments, which are part of the biological properties that realize normal vision. They take this to be evidence that, in scientific practice, types at the psychological level are not always altered or divided after heterogeneity in the lower-level properties that perform them is discovered. This suggests that the psychological level of explanation can be autonomous, and that Kim was not right (and that if he was, this would have to be decided empirically, not a priori).[64] If Aizawa and Gillet are right, and multiple realization is a real phenomenon[65], it is, I think, sufficient to deny the identity between mental and neural types, and consequently to deny the reduction of psychology to neuroscience.

---

[62] As the authors observe, even in the case of memory, differences at the psychological level between these two types of memory also influenced the decision to separate the two types. That is to say, even in this case the decision to deny multiple realization, and to split memory into two types, was not only due to what the neuroscientific theories had to say about the brain, but also what seemed appropriate from the point of view of psychological theories.

[63] According to them, "One finds that in actual scientific practice not all discoveries about differences in realizing properties influence higher level theory in the same way. In particular, scientists do not uniformly adopt the eliminate-and-split strategy. (…) Psychological theory shapes how psychology accommodates the discovery of differences in neuroscientific realizers in partnership with lower level theories, rather than the lower level theories simply necessarily dictating changes through their discoveries." (AIZAWA; GILLETT, 2011, p. 205).

[64] But, as Aizawa and Gillet note, "this is not to say, however, that scientists simply dismiss differences in lower level realizers as irrelevant to the higher level theory or properties. There is not *that* kind of autonomy of psychology. Scientists often study differences in lower level realizers as a means of explaining what they refer to as individual differences, differences from one human to the next." (2011, p. 214). In this respect, Aizawa and Gillet seem to have a more reasonable view than that of Fodor, for whom psychology has a complete autonomy with respect to neuroscience.

[65] This is a point about which there is a lot of discussion (see BICKLE, 2013), and it deserves a more detailed treatment. It could be questioned, for example, how types of brain states are to be individuated. It is possible that

However, although multiple realization is sufficient to stop the reduction of psychology in the present scenario, I think that the use of this argument would not be sufficient in all scenarios. We have seen that both Fodor and Putnam think that it is not possible to identify mental types to neural types, because it is a logical possibility that beings with a non-biological constitution have mental states. In "Special sciences" Fodor in fact takes this to be a nomological possibility. According to him, even if there is a coextension between mental and cerebral types, they cannot be identical because "it seems increasingly likely that there are nomologically possible systems other than organisms (viz., automata) which satisfy the kind predicates of psychology but which satisfy no neurological predicates at all." (FODOR, 1974, p. 18). However, the truth is that we do not know whether or not it is a nomological possibility that robots can have intentional states. If we found out that types of mental states are in fact coextensive with types of neural states, and that there seems to be something in the biological material of the brain necessary for mentality (thus discarding the nomological possibility of automata with mental states), there would be nomologically necessary bridge laws connecting both types. In this scenario, the mere logical possibility of the occurrence of mental states in non-biological systems would not be a good reason to deny the reduction of mental types to neural types. Scientific laws need not take into account, and in general do not, all the logical possibilities of the phenomenon they are dealing with. It is too strong a requirement for reduction that the two types involved should be identical in all possible worlds.

My point, then, is that supposing neuroscience observed that types of mental states always correspond to the same types of brain states (and that we somehow knew that non-biological beings cannot have psychological states), the mere logical possibility of multiple realization would hardly be a good reason to stop the reduction of psychological types to neural types. It is more likely that in the scenario we are imagining, this coextension would be taken as a sign that there is an identity between mental and neural types. I think then that multiple realization, if indeed proven, is sufficient to deny the identity between mental types and brain types. In the current scenario, there is reason to believe that tokens of the same type of mental state can be realized by tokens of brain states belonging to different types. But if new research were to show that there is

---

brain states that perform the same mental type, at first considered heterogeneous, are in fact not heterogeneous; they are only perceived as heterogeneous because we still do not know how to properly individuate brain states. If that were the case, multiple realization would be a phenomenon that is only apparent. It could be revealed false if we had more knowledge of how the brain works. While these are interesting questions, for our purposes, a superficial idea of what is at stake is sufficient.

in fact a one-to-one correspondence between mental types and neural types, the mere logical possibility of multiple realization, and of automata with mental states, would hardly be enough to convince the identity theorist that there is no identity between mental and neural types, and to stop the reduction of psychology to neuroscience.

At this point we can imagine Kripke's intervention. In *Naming and Necessity* (1980), Kripke criticizes the identity theory based on the idea that identity is a relation such that, when it holds, it holds necessarily. According to him, the relationship between pain and activated C-fibers (or whatever the neural correlate of pain is) does not hold with necessity. This is because one can conceive that there is the sensation of pain without there being activated C-fibers, or that there are C-fibers activated without the sensation of pain. From the fact that one can conceive one property occurring without the other, Kripke concludes that it is possible that there is pain without there being activated C-fibers, and vice-versa. Given that if pain and C-fibers firing were identical, the identity would be necessary, and that it is not necessary for the sensation of pain to be always accompanied by activated C-fibers (because the opposite is conceivable and therefore possible), Kripke concludes that the relation that pain maintains with C-fibers is not a relation of identity. Kripke would then say that even if pain is always associated with activated C-fibers, these two properties are not identical and, possibly, that psychology cannot be reduced to neuroscience.

Smart and Place, proponents of the identity theory, could counter that they never considered the identity they talk about to be necessary. Smart states, for example, that

> there can be contingent statements of the form "A is identical with B," and a person may well know that something is an A without knowing that it is a B. An illiterate peasant might well be able to talk about his sensations without knowing about his brain processes, just as he can talk about lightning though he knows nothing of electricity. (SMART, 1959, p. 147).

Smart and Place believe that the relation of identity that holds between mental and cerebral states is a scientific identity like any other, similar to the identity between lightning and electric discharge, or between heat and molecular kinetic energy, and they consider that all these identities are contingent. By "contingent" they seem to mean both that these relations of identity

might not have been the case, and that they are truths that must be discovered empirically (as opposed to analytical truths).

Kripke would certainly agree that scientific identities are usually known *a posteriori* (by empirical investigation). But against Smart and Place, he argues that the supposed identity between mental properties and neural properties is not as analogous to other scientific identities as it might appear at first. According to him, it is in fact not conceivable that heat is anything other than molecular kinetic energy (although it is conceivable, and therefore possible, that the *sensation* of heat is not accompanied by molecular kinetic energy). In the case of pain, however, the sensation of pain *is* the pain, and though pain cannot be conceived without the sensation of pain, it can be conceived as occurring without the activated C-fibers. And, to the extent that the relationship between pain and activated C-fibers does not hold with necessity, pain is not identical to whatever its physical correlate is.[66]

However, what Kripke's argument shows is at most that it is *logically* possible for pain to exist without activated C-fibers, and vice versa. So Kripke, like Fodor, denies the identity between mental and brain types based on a logical possibility. But his argument does not show that it is nomologically or physically possible that one exists without the other. So Smart and Place could still say that what interests them is to argue for an identity between mental and brain properties that is only nomologicaly necessary. Someone who wishes to promote the reduction of psychology to neuroscience might consider that all that it actually requires is the nomological identity between mental and cerebral types.

A second alternative for the identity theorist would be to counter Kripke's argument by denying that conceivability implies possibility. Kripke assumes that it is possible for pain to occur without a specific physical correlate because he can conceive this. If this inference was somehow barred, and assuming mental types are coextensive with physical types, the identity theorist could still maintain that the relation there is one of identity.

A third alternative for a reductionist would be to accept that Kripke is right that mental states are not identical to brain states, and that "nomological identity" is an oxymoron (because identities are always necessary), but still refuse to accept premise (1) of Fodor's argument (the assumption that reducing psychology to neuroscience requires that mental properties maintain a

[66] Searle says, about Kripke's argument, that "it is essentially the commonsense objection in a sophisticated guise. The commonsense objection to any identity theory is that you can't identify anything mental with anything nonmental, without leaving out the mental." (SEARLE, 1992, p. 39).

relation of identity to neural properties). For the purpose of reduction, it could suffice that every mental type corresponds to a certain cerebral type with nomological necessity. Coextensitivy might be all we need for reduction.

2.2 On the necessity of multiple realization for the autonomy of psychology

My point so far has been simply that, if science reveals that mental types are multiply realizable, then this is enough to deny that psychology reduces to neuroscience. If, however we found out that mental types nomologically correspond to brain types, appealing to the mere logical possibility of multiple realization (with Fodor), or of mental states occurring without corresponding brain states (with Kripke) would not be very convincing against someone who wants to defend the reduction of psychology to neuroscience. I believe, however, that there are other ways of arguing against the reduction of psychology to neuroscience. As we have seen, multiple realization serves mainly to deny the identity between mental and brain types. In denying this identity, we deny the reduction of psychology to neuroscience. But do we *need* multiple realization to deny this identity and guarantee the autonomy of psychology? And do we even need to deny the identity in order to preserve the autonomy of psychology?

Let us consider again the scenario in which we've got a complete neuroscience and, contrary to what seems to be true, it has been found that psychological types reliably correspond to specific neurological types. Assuming that it is false that tokens of the same mental type can be physically realized in tokens of heterogeneous physical types, i.e. assuming that multiple realization is false, should we conclude that neural types are identical to mental types, and therefore that psychology can be reduced to neuroscience? Do we really need multiple realization to deny identity between types and block the reduction? In addition, do we really need to deny that these types are identical to preserve the autonomy of psychology?

I believe that some kind of anti-reductionism would hold up even if, in a complete neuroscience, we found that mental types and brain types are coextensive. A natural anti-reductionist approach would be to say that mental types would not be identical or reducible to brain types, even if they were coextensive, because psychological types capture different properties from the properties that are captured by neuroscience. The tomato I'm imagining now

is red, but the brain state correlated with it involves, say, 10.000 neurons connected in a particular way. If we want to, we can say that we have here one event, which admits two descriptions: one given by psychology (the redness of the imagined tomato) and one given by neuroscience (the number of neurons connected and activated), each highlighting different aspects. Even if these two properties always occurred together, they would still seem to be different. This would be a form of property dualism. We could accept that neuroscience gives us an explanation of the biological constitution of mental states, but we do not have to accept that mental and brain types are identical, even if they were coextensive. These properties also play different roles in our explanatory practices. When we talk about the causes of behavior, for example, we typically refer to mental properties. We want to draw attention to the qualitative or intentional aspect of mental states, which are not highlighted when we talk of neural structures, for example. In general, we consider that a mental property is causally responsible for a behavior, regardless of what its underlying physical property is (I think it was my pain that caused my desire to take an aspirin, not my brain state, even assuming that pain always consists of the same physical properties). Substituting neural predicates for mental predicates in explanations of behavior would render them not only unintelligible, but beside the point.

The fact that psychology highlights aspects not highlighted by neuroscience could then be used to deny that neural types are identical to mental types, without the need to assume multiple realization. Psychology could still maintain its autonomy, even if its types were nomologically coextensive with neural types. That is, even if tokens of the same mental type were always realized by tokens of the same neural type, psychology could still claim a vocabulary and explanations distinct from those of neuroscience, without becoming dispensable. One possible view is that mental properties supervene on physical properties. In this view, one could accept that there are two properties, one mental and one physical, where the mental depends on the physical to exist, but they would still not be identical, because they highlight different aspects of reality. This view is still compatible with token physicalism, understood as the view that to each occurrence of a mental state corresponds an occurrence of entities that physics deals with (and therefore mental properties do not belong to an immaterial substance that is outside the scope of scientific investigation).[67] According to this approach, it would be possible to adopt an anti-

[67] The issues here are in fact more complicated than that. If we understand token physicalism as the idea that each mental state token is *identical* to a brain state token, then it seems that it is not exactly compatible with property dualism. We might understand token physicalism as saying that my tokening of, say, pain, is the same event as my

reductionist stance even in a scenario where multiple realization is false, which suggests that multiple realization is not necessary to preserve the autonomy of psychology.

Perhaps, however, we need to qualify the kind of anti-reductionism we would be talking about in this scenario. In response to my observations so far, the identity theorist might say that what we would have in this scenario would be only an *epistemological* autonomy of psychology from neurology, but not an *ontological* autonomy. That is, one could deny that the fact that psychology and neuroscience approach different aspects implies the falsity of the identity theory, from an ontological point of view. According to this critical view, if every token of mental type X necessarily corresponded to tokens of brain type Y, type X would be ontologically identical to type Y, even if from an epistemological point of view, X could not be reduced to Y. Mental states are nothing more than brain states. If this scenario were true, the identity theory would be true, and psychology would be ontologically reduced to neuroscience, even if its epistemic autonomy were preserved. Headaches would simply be neural states of type Y, even if, when we talk about headaches, we highlight one aspect of these events that is not highlighted by the expression "neural state Y."

I think the assumption that mental state types would be identical to neural types would be unwarranted, because it still seems intuitive to assume that psychological descriptions of mental events refer to properties which brain states, qua brain states, don't have (being true or false, in the case of a belief, or intense, in the case of pain). But assuming we accept the view of this imaginary supporter of the identity theory, I think that the epistemological autonomy of psychology is no less important than the ontological autonomy. Fodor also recognizes that psychology has an epistemological autonomy with respect to neuroscience. According to him,

> it is often the case that *whether* the physical descriptions of the events subsumed by such generalizations have anything in common is, in an obvious sense, entirely irrelevant to the truth of the generalizations, or to their interestingness, or to their degree of confirmation, or, indeed, to any of their epistemologically important properties. (FODOR, 1974, p. 15).

---

tokening of C-fibers firing – it is just one event that admits two different descriptions. Other tokens of pain might be the same event as the tokening of property WZX. But we could then ask what is it about these two tokenings of properties that makes them tokenings of the same event. How do we individuate events? But I won't pursue these questions any further.

In *Psychosemantics*, he notes that

> Even if psychology were dispensable *in principle*, that would be no argument for dispensing with it. (Perhaps geology is dispensable in principle; every river is a physical object after all. Would that be a reason for supposing that rivers aren't a natural kind? Or that 'meandering rivers erode their outside banks' is untrue?) What's relevant to whether commonsense psychology is worth defending is its dispensability *in fact*. And here the situation is absolutely clear. We have no idea of how to explain ourselves to ourselves except in a vocabulary which is *saturated* with belief/desire psychology. (FODOR, 1987, p. 09).

However, at least in "Special sciences," Fodor clearly wants to preserve both types of autonomy. I think Fodor insists on the multiple realization argument precisely because it guarantees the ontological autonomy of psychology in relation to neurology.[68] If multiple realization is true, it shows that there is something in the way the world is organized that guarantees the autonomous status of psychology; that this autonomy is not just a matter of our knowledge of mental states. As Fodor says,

> I am suggesting, roughly, that there are special sciences not because of the nature of our epistemic relation to the world, but because of the way the world is put together: not all the kinds (not all the classes of things and events about which there are important, counterfactual supporting generalizations to make) are, or correspond to, physical kinds. (FODOR, 1974, p. 24).

If multiple realization is true, psychology gains an ontological autonomy, because psychological types will have a form of existence which is independent of specific neural types. Multiple realization is a good way of securing the ontological autonomy of psychology. If we accept that it is likely that mental types are indeed realizable in different neural types, the types of psychology do indeed have some kind of independent existence of specific types of neuroscience. I think, however, that epistemological autonomy also guarantees a respectable status for

[68] However, in his latest book, Fodor seems happy with the epistemic autonomy of psychology: "though we assume that the mechanisms by which mental causation is implemented are in all likelihood neural, we don't at all suppose that psychological explanations can be reduced to, or replaced by, explanations in brain science—no more than we suppose that geological explanations can be reduced to, or replaced by, explanations in quantum mechanics. Confusions of ontological issues about *what mental phenomena are* with epistemological issues about *how mental phenomena are to be explained* have plagued interdisciplinary discussions of how—or whether— psychology is to be "grounded" in brain science." (FODOR; PYLYSHYN, 2015, p. 02).

psychology; psychology as a science is not dispensable because it searches for regularities and explanations of mental states as we know and experience them, and not as something else. Psychology makes generalizations involving mental states, independently of reference to brain processes, and we have every reason to assume that things will continue to be so, even assuming that the identity theory proves true. RTM and CTM could still be true, even if the identity theory were true.

This may be in conflict with what I said earlier in this chapter, that is, that the identity theory is a theory that competes with RTM. But in fact both could be true, as long as we softened the ambitions of each one of them, and we established with precision the level of explanation that each one is supposed to capture. In fact, maybe none of them captures the unique essence of mental states; RTM and CTM say what the psychological level of explanation consists of, and the identity theory says that there are bridge laws between types of psychology and types of neuroscience, about which RTM has nothing to say. Psychological theories can legitimately take mental states and processes as their object of inquiry, even if they are ontologically reducible to entities of more a basic science. If the role of psychology is to investigate mental states *qua* representational states and mental processes *qua* computational processes, its autonomy is guaranteed, at the very least from an epistemological point of view.[69] I think therefore that neither multiple realization, nor the falsity of the identity theory, are necessary to guarantee the legitimacy of a computational and representational characterization of mental states and processes, which is an intermediate level of description, in between the level of common-sense beliefs and desires, and the purely biological level – even if, without multiple realization, we may be reticent about the ontological independence of the natural types of psychology.

Psychology thus seems to have at least an autonomous epistemic status with regard to neuroscience. The same holds true for common sense psychology. To make this point more vivid, it may help to recall the obvious, which is that explanations of behavior based on mental states are indispensable for our understanding of ourselves. As Fodor says, "commonsense psychology works so well it disappears" (1987, p. 03). We are beings with beliefs, desires, hunches, fears, ambitions. We think of the emptiness of existence, we aspire to a good career, etc. We feel pain, hunger, heat, anger and tiredness. We perceive the world through the senses. We see objects,

[69] Fodor notes that "there seems to be a level of abstraction at which the generalizations of psychology are most naturally pitched. This level of abstraction cuts across differences in the physical composition of the systems to which psychological generalizations apply. In the cognitive sciences, at least, the natural domain for psychological theorizing seems to be all systems that process information." (FODOR, 1981a, p. 118).

which appear to be colored or pale, near or distant, large or small. We hear sounds. We feel sweet or bitter tastes, pleasant and unpleasant smells. We act driven by mental states like these, and we explain and predict the behaviors of ourselves and others by appeal to mental states like these.

When we seek an explanation of why so-and-so acted in a certain way, what we want to know is usually: what mental states led to that action. If someone asks me why I spent 250 reais on a concert ticket, an acceptable explanation would be that I really like the band and that I believe I am going to have a good time. To say that I bought the ticket because my brain was in state X and not Y would be beside the point (although it may be true that, if my brain were not in the state it was, I would not have bought the ticket). The same goes for the behavior of animals, or of mythological and fictional characters. The best way to explain why Miguelinho, my cat, is looking at me and meowing, next to his bowl full of a new kind of food, is that he is not happy with his new food and he wants me to give him the old one (when offered the old food, he simply eats it, no complaints). If I were asked why Anna Karenina killed herself, no satisfactory explanation would say that it was because she had been having brain states of the types WJD and TPG. The reason this explanation would be unsatisfactory is certainly not that fictional characters, strictly speaking, do not have brains.

This does not mean that reference to something that goes on in the brain is always irrelevant in explaining behavior. Take the case of Charles Whitman. He was a normal person until, in 1966, he killed his mother and his wife, and then shot and killed several people at the University of Texas. In his suicide note, he indicated that he was having irrational thoughts, and asked that his body be examined after his death. At the autopsy, they found a brain tumor close to the amygdala, which is an area associated with emotions. So it is reasonable to assume that the brain tumor at least partly explains his abrupt change in behavior and its tragic consequences. I say "partly" because, even in this case, the tumor is only relevant to explain why Whitman killed several people if we assume that it affected his ability to control anger, for instance – which is a mental state. Without the supposition that the tumor interfered with relevant mental states, mentioning the tumor would also be irrelevant to explain his actions.

What makes reference to the brain, in the typical case, irrelevant to an adequate explanation of behavior? I do not think I have a good answer to this question. The most I can say is that we are so made that we have all sorts of mental states, as mentioned above, through which we know ourselves and the world. By introspection, I may note that I have beliefs, sensations, etc.,

and that these states appear to cause my behaviors. Other people, similarly, behave the way they do due to similar mental states. What happens in my brain and body, under a biological description, is something totally foreign to me. I feel pain, or I have the urge to eat ice cream. But I do not know what goes on in my brain when I have these states. The point is not simply that we have a limited knowledge of our bodies.

Suppose I knew everything that goes on in my body when I have any mental state. And let us suppose that all other human beings possess the same level of knowledge. Would it be any less irrelevant to explain someone's behavior by appealing to their brain states instead of their mental states? I suspect that knowing everything that goes on in the brain and body would not make us think that the sensation of pain, or the desire to eat ice cream, has a less important causal role in the explanation of why I screamed, or why I went to the ice cream shop. This is because mental states would continue to appear to us as real and causally effective as they appear now, and indispensable to our understanding of ourselves and others. Again, we are beings who have these mental states, and who know the world through them. This is how we are made. Knowing the states of the brain and body would not substantially change our explanatory practices, because we would continue to feel pain, see red, and have the desire to eat ice cream. Even if we always knew what states of the brain or body corresponded to each of our mental states, mental properties would still appear to us as different from bodily properties.

2.3 Weak vs. strong multiple realization

I argued in the previous section that multiple realization is not necessary for the explanatory autonomy of psychology. Psychology would still be autonomous even if multiple realization were false. Either because we could falsify the identity theory by other means (by accepting property dualism), or because, even if accepting the identity, mental properties would still have an epistemic autonomy in relation to brain properties. But, if it is indeed true that there is no one-to-one correspondence between psychological types and neural types, this suggests that psychological types are ontologically independent of specific neural types.

However, multiple *neural* realization, the realizability of the same mental state type in different types of brain states (which we can call "weak multiple realization"), does not guarantee that psychological types are independent of *any* neural type; it does not show that psychological

types are unrestrictedly realizable, even in non-biological systems (a view we can call "strong multiple realization"). We have seen that in addition to supposing that psychological types are independent of specific neural types, Fodor, Putnam, and other functionalist philosophers tend to ascribe to psychological states this stronger type of independence, supposing that biological material is not necessary for their existence. In several writings, Fodor seems to infer the possibility of strong multiple realization (which includes nonbiological materials) from the assumption that the functional and computational explanations of mental states give us their complete essence. That is, to the extent that Fodor thinks that propositional attitudes are nothing more than functional relations to a symbol, and that mental processes are nothing more than transformations of these symbols, he seems to believe that nothing prevents intentional psychological states from existing in any kind of material. Basically, Fodor seems to believe that functional and computational states are ipso facto autonomous and multiply realizable, insofar as the material of which they are made is not specified at this level of description. Recall that Fodor states that "the psychological constitution of a system seems to depend not on its hardware, or physical composition, but on its software, or program" (FODOR, 1981a, p. 117). Later, his idea that functional properties are always multiply realizable becomes apparent, when he says: "Pace Kim, being jade is not relevantly like having a *functional (i.e., MR) property*. (…) There's a difference between being a *functional property (being multiply realized)* and being a disjunctive property" (FODOR, 1997, p. 14, my emphasis).[70] Ned Block seems to hold the same view:

> According to cognitive science, the essence of the mental is computational, and any computational state is 'multiply realizable' by physiological or electronic states that are not identical with one another, and so content cannot be identified with any one of them. (BLOCK, "Can the Mind Change the World?", apud Kim).

Fodor does accept that "it is reasonable to expect that our minds may prove to be neurophysiological (token physicalism for human beings)" (1981a, p. 120). But even if this is the

[70] In his latest book with Pylyshyn, Fodor seems to hold a less radical view on multiple realization: "it is convenient to suppose that mental representations are neural entities of some sort, but a naturalist doesn't have to assume that if he doesn't want to. We are officially neutral; all we insist on is that, whatever else they are or aren't, mental representations are the sorts of things whose ultimate parts basic science talks about. There may be a better way out of the puzzle about how mental contents can have causal roles, but we don't know of any." (FODOR; PYLYSHYN, 2015, p. 07).

case for humans, he doesn't think that mentality in general depends on neural states to exist, for he thinks that robots with intentional states are a nomological possibility.

However, it is one thing to say that representational functionalism has the advantage of preserving this possibility, given that "the software description of the mind does not logically require neurons." (FODOR, 1981a, p. 120). It is another to infer strong multiple realizability (that intentional states are independent of *any* neural state) from the idea that we can characterize intentional states without reference to the physical constitution of the system that has them. The first idea seems reasonable, but the second does not. Functional characterization does not imply multiple realization. In my view, this inference is only valid if we accept that the functional and computational characterizations capture the complete essence of intentional states, which is questionable. Even if it is true that the biological material of a system is irrelevant to the existence of intentional states, we still do not know whether this is true or not. We cannot simply stipulate that intentional states can (nomologically) exist in any material because a psychological description of these states need not refer to any brain state. The most we can say is that psychology captures a functional, or computational, level of explanation of mental states that does not refer to brain states. But the mere existence of this level is not enough to conclude that all cerebral states are inessential to mental states. Even if psychology is independent of neurology from an epistemological or explanatory point of view, it is still possible that there is something in the material of the brain necessary for intentionality, and therefore that the instantiation of the right computational processes is not sufficient for intentional mental states (this topic will reappear in section 3).

Thus, neither epistemological autonomy nor what we might call weak ontological autonomy (the idea that mental types occur in different neural types) – a consequence of weak multiple realization – guarantee that psychology has a strong ontological autonomy with respect to neuroscience – according to which mental states are strongly multiply realizable. It is one thing to say that there are two levels of description that apply to mental states, and that the description of the functional and computational organization of these states occurs essentially without reference to the description of their neural constitution. It is another thing to say that the essence of these mental states is entirely captured by their functions and computations, and that their neural constitution is irrelevant not only to the description of the psychological level but also to

the very existence of mental states.[71] Of course we can stipulate that mental states are essentially characterized by their computations, and entirely independent of a biological constitution. But at the current stage of mind and brain knowledge, I think this would be an unwarranted assumption. As far as we know, it is possible that there is something in the biological constitution of an organism that is essential for the existence of mental states; it is possible that as much as we could reproduce the computational and functional organization of an organism artificially, we would not obtain genuine mental states. Basically, we do not know what the essence of mental states is, and we cannot rule out the hypothesis that they need a biological basis.

3. Searle's challenges

We have so far seen some aspects of behaviorism and the identity theory, and how, unlike RTM and CTM, they don't vindicate common sense belief/desire psychology, to the extent that they reduce mental states to other types of entities. Here I'm going to consider some of Searle's objections to ideas that are central to cognitive science (and to CTM).

As we have seen in the first chapter, Fodor suggests that we should conceive mental processes as computational processes, for he thinks it can help us understand how at least some mental processes (like some rational processes) can occur in a mechanistic, or naturalistic way. According to him, "computers show us how to connect semantical with causal properties for *symbols*. So, if having a propositional attitude involves tokening a symbol, then we can get some leverage on connecting semantical properties with causal ones for *thoughts*" (1987, p. 18). The idea is that if we suppose that intentional states involve the instantiation of a symbol, we can assume that the syntax of this symbol plays the mediating role between its semantic content and its causal power. The computer manipulates symbols only by virtue of their syntactic forms, which are ultimately physical, and not by virtue of the semantic contents of those symbols. As Fodor says, "the syntax of a symbol might determine the causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open" (1987, p. 18). If we treat mental processes as computational processes, i.e., as transformations of symbols by virtue of their

[71] Fodor also seems to think that understanding how the brain works doesn't help us understand how the mind works (this is suggested in FODOR, 1999). But this is a central assumption in cognitive neuroscience, and if we are to follow Fodor's own recommendations, we should take the assumptions of scientists seriously. The study of the brain can illuminate the study if the mind, and vice-versa, without one being reducible to the other in any way.

syntactic forms, we can understand how mental processes occur mechanically while preserving semantic relations. Fodor is committed, at least for a particular class of mental states and processes, to the existence of a level of computational explanation that would be a level between common-sense psychology and neuroscience. This intermediate level would consist of mental representations, conceived as symbols with constituent syntactic structure and compositional semantics. The idea then is that there is a language of thought, and that mental processes are computational transformations of symbols or representations of that language.

John Searle, one of the critics of the computational theory of the mind, formulates several problems against the idea that there is a computational level of explanation of mental phenomena. In this section, I will introduce and discuss some of these. One is derived from his famous Chinese room argument, formulated in the 1980s. Searle offers this argument to show mainly that programs are not sufficient for intentionality. This has at least two consequences. First, that mental processes cannot be characterized only as computational processes, since computational processes are merely manipulations of meaningless symbols, whereas mental states and processes usually involve intentionality. Second, that computers can never have intentional states just by virtue of the implementation of a program. It is important to note that Searle directs his argument especially against proponents of what he calls strong Artificial Intelligence, according to which properly constructed programs are sufficient for the existence of intentional states (that is, states that represent or are about something).

Searle also raises problems against the idea that the study of brain is irrelevant to the study of the mind, since he holds that something with the causal powers of the brain is necessary for intentionality. In the 1990s, Searle began to make broader criticisms of cognitive science. He goes on to observe, basically, not only that computational processes are not sufficient for there to be intentionality, but that it is not even necessary to suppose that mental processes involve computational processes. According to him, there is no need to suppose that there is a language of thought, nor any other mental states or processes that are inaccessible to consciousness. There are only two levels of explanation of mentality: common-sense psychology and neurobiology.

We will see that Fodor is not a strong AI advocate, but he thinks that (at least some) mental processes can be characterized as computational processes. As an advocate, to some extent, of functionalism, Fodor also believes that computers, at least in principle, could have intentional states. We have also seen that Fodor believes that psychology is an autonomous

science, which may suggest the idea that, for him, the study of the brain is not particularly relevant to the study of how the mind works. In what follows, I will first present the Chinese room argument, followed by some of the other problems Searle raises against cognitive science. In part 3.2, I'll show how Fodor reacts, or could react, to each of Searle's central points. Finally, in 3.3, I indicate who I think is right about what.

3.1 Chinese room, brains and syntax

Let us briefly review the Chinese room thought experiment, first formulated by Searle in "Minds, Brains, and Programs" (1980). Searle, who does not know Chinese, imagines himself locked in a room, where he receives two stacks of Chinese writing, and a set of rules in English, showing how to relate the first stack of texts to the second stack. He then receives a third pile of Chinese writing, accompanied by rules in English showing how to associate this third stack with the first two, and with instructions to send back other sets of symbols in Chinese. Suppose these rules are of the type: if you get the symbols X Y Z, send back the symbols you find in line 3 of book 2. Searle then sends back H K L. When given a certain sequence of symbols in Chinese, Searle follows the rules given in English, and sends back other sequences of symbols out of the room. One can imagine that after a while Searle becomes so efficient at this task that he can memorize the rules and the symbols contained in the stacks, so that he no longer needs to consult them to follow the instructions.

Unbeknownst to Searle, the people in charge outside the room call the first stack "Script" and the second "story". The third stack has questions about the story; the rules followed by Searle are called "program", and the symbols he returns are the answers. What someone outside the room, who knows Chinese, sees is that from inside the room come answers in Chinese to questions asked about a story. And these answers are indistinguishable from those that a Chinese speaker could give. The answers given by Searle are therefore compatible with the answers that a Chinese speaker could give if asked the same questions about this story. But though Searle somehow behaves like a Chinese speaker, who understands the story, he does not understand the story, and has no idea what his answers mean. Searle does not know Chinese, and does not even know he's answering questions about a story. He only "produce[s] the answers by manipulating

uninterpreted formal symbols" (SEARLE, 1980b, p. 418). In other words, he implements a program, manipulating symbols based merely on their syntactic forms, with no knowledge of the semantic content of the symbols.

With the Chinese room argument, Searle intends, first of all, to overturn one of the central ideas of what he calls strong AI, which is that the mind is nothing more than a computer program. According to strong AI followers, a properly constructed program, regardless of how it is implemented, is sufficient for the literal existence of cognitive or intentional states. Searle's point is that a computer, however sophisticated it may be, when following a program, does nothing more than he does inside the room. The computer instantiates a program that indicates which outputs should be given when it receives such and such inputs. The symbols it transforms are, from the point of view of the computer itself, meaningless. Just as Searle does not understand Chinese, although he follows a program for manipulating symbols like a Chinese speaker, neither does a computer understand the symbols it manipulates. It just follows a program. Thus, Searle intends to show that the instantiation of a program is not a sufficient condition for the existence of intentional states (such as states related to the comprehension of a language). Searle reformulated that point ten years later at a conference entitled "Is the brain a digital computer?":

> Since programs are defined purely formally or syntactically and since minds have an intrinsic mental content, it follows immediately that the program by itself cannot constitute the mind. The formal syntax of the program does not by itself guarantee the presence of mental contents. (…) The argument rests on the simple logical truth that syntax is not the same as, nor is it by itself sufficient for, semantics. (SEARLE, 1990a, p. 21).

Thus, Searle's first point is that mental processes cannot be purely and simply computational processes, because mental processes involve states with intrinsic semantic content, whereas computational processes are formal processes that do not require symbols with intrinsic semantic content. A program by itself cannot be what explains the semantic aspect of mental states, because programs can exist without any intentional state. Characterizing the mind as a computer program is fated to be at best an incomplete characterization.

The second point is that, if all computers do is to implement a program, then computers cannot have cognitive states; they cannot think, or have intentional states, because programs are not sufficient for the existence of intentional states. Even if we were to reproduce the program

allegedly implemented by the brain, implementing that program on a computer will never be enough to give it intentional states, because programs are formal, and mental states are intentional. A third consequence of Searle's argument is that behavioral criteria for assigning mental states to computers, such as the Turing test, are not adequate. According to the test devised by Turing (1950), a machine will have intelligence if it is capable of deceiving an evaluator, in a blind test, on whether it is a machine or not. So the appearance of the machine, and the material of which it is made of, are taken to be irrelevant for its possession of cognitive states. What is relevant is whether it follows a program that allows it to behave like a human being. According to Searle, the Turing test does not really determine whether a computer is capable of thinking. This is because even if a machine can behave humanly enough to pass the test, all it does is to instantiate a program that formally manipulates symbols that have no intrinsic semantic content, so it doesn't really have intentionality.

There are a number of objections that can be made against Searle's argument, and he himself considers some of them in his 1980 paper. One is the robot's reply, which accepts that Searle, locked in a room, certainly does not understand Chinese. The same would occur with a computer that does not receive proper input from the environment. However, if we imagine a robot capable of interacting causally with the environment, i.e. capable of receiving inputs from the environment and of producing outputs, it is not clear that it would be incapable of having intentional states. For Searle, however, interaction with the environment does not add meaning to the manipulated symbols. According to him, this situation is not very different from that of the Chinese room. Searle suggests that we imagine that he, instead of a computer, is inside a robot. His intuition is that he will continue to manipulate symbols that have no meaning for him:

> the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols. (SEARLE, 1980b, p. 420).

So Searle thinks that the robot's interaction with the world does not give it intentional states, because it continues to only follow a program. A robot continues to manipulate meaningless symbols. It only differs from an ordinary computer in that it receives symbols from a different source, and in that it can send symbols that produce motor responses.

But how then do we understand stories in our mother tongue, for example, if it is not because we instantiate programs, or because we are in appropriate causal relations with the world? What guarantees that we have intentionality? What allows a Chinese speaker, and not a computer, to know the meanings of Chinese words and to be able to think about the world? According to Searle, what explains the intentionality in us is the fact that we are

> a certain sort of organism with a certain biological (i.e. chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality. (SEARLE, 1980b, p. 422).

So for Searle, intentional states depend on something with similar causal powers of the brain to exist. I think it is important to note that Searle does not mean that intentionality can only arise from something that has exactly the same physico-chemical composition of the brain. Although it is common to interpret him this way, what he means is somewhat trivial. His idea is that if we accept that intentionality is caused by the brain, we must also accept the trivial idea that, in order for there to be intentionality, there must be something that has the power to cause that intentionality, just like the brain is capable of causing it in us.[72] This observation is somewhat trivial because one would hardly deny that intentionality depends to some extent on the existence of some physical substrate (disregarding the dualists). But this observation falls within the context of his critique of strong AI, whose proponents would say that intentionality is a product only of the implementation of a program, and that the physical realization of that program is irrelevant for the existence of intentionality. Searle, on the contrary, points out that there will only be intentionality if there is something physical that is capable of producing that intentionality – and presumably not everything has that power. The brain is one thing (and, as far as we know, the

[72] In fact, Searle himself recognizes that his observation is trivial: "one thing we know before we even begin the investigation is that *any system capable of causing consciousness must be capable of duplicating the causal powers of the brain.* If, for example, it is done with silicon chips instead of neurons, it must be because the chemistry of the silicon chips is capable of duplicating the specific causal powers of neurons to cause consciousness. It is a trivial logical consequence of the fact that brains cause consciousness that any other system capable of causing consciousness, but using completely different mechanisms, would have to have at least the equivalent power of brains to do it. (Compare: airplanes don't have to have feathers to fly, but they do have to share with birds the causal capacity to overcome the force of gravity in the earth's atmosphere.)" (SEARLE, 1992, p. 92).

only) that has that power. But Searle also accepts that there is nothing that prevents a priori the possibility of other types of physical substrates giving rise to intentionality. According to him,

> Perhaps other physical and chemical processes could produce exactly these effects [of intentionality]; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. That is an empirical question, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll. (SEARLE, 1980b, p. 422).

> I have not tried to show that only biologically based systems like our brains can think. Right now those are the only systems we know for a fact can think, but we might find other systems in the universe that can produce conscious thoughts, and we might even come to be able to create thinking systems artificially. I regard this issue as up for grabs. (SEARLE, 1990b, p. 27).

> The PDP-10 is powered by electricity and perhaps its electrical properties can reproduce some of the actual causal powers of the electrochemical features of the brain in producing mental states. We certainly couldn't rule out that eventuality a priori. But remember: the thesis of strong AI is that the mind is "independent of *any* particular embodiment" because the mind is just a program and the program can be run on a computer made of anything whatever provided it is stable enough and complex enough to carry the program. The actual physical computer could be an ant colony (one of their examples), a collection of beer cans, streams of toilet paper with small stones placed on the squares, men sitting on high stools with green eye shades—anything you like. (SEARLE, 1982).

Searle is not saying, therefore, that for a creature to have mental states, it needs to have a brain with the same composition as ours. For him it is an open question whether something different from the brain could give rise to mental states. His point is that it is the brain which causally explains the presence of intentional states in us, not the instantiation of some program. Mental states such as pains, thoughts, memories and sensations, are caused by specific neurobiological processes. Formal aspects of these biological processes are not causally responsible for the existence of these mental states. In the same way, an artificially constructed machine could in theory think, but not because of the implementation of a program, but rather because its hardware would have the same causal powers as the brain. The intentionality of the machine would not be a product of the instantiation of a program, since programs are never sufficient for intentionality. It would be a product of the physical organization of the machine. In Searle's view, computers can never be cognitive beings by virtue of the implementation of a program; they can never think if they lack a hardware that has the same causal powers as the

brain. A purely formal program cannot give rise to intentionality as the neurobiological processes of the brain can.

It is interesting to see that Searle's position is not entirely incompatible with the typically functionalist idea of multiple realization, since Searle admits the possibility that both artificial machines and beings with a physicochemical structure different from ours have mental states. The difference seems to be that, for him, the hardware plays a much more important role than for a functionalist. Functionalists usually accept that beings of different physical organizations could have cognitive states as long as their functional organization is similar to ours, and a similar functional organization could in theory be implemented in any kind of material. Searle, on the contrary, certainly denies that mental states can be realized in any kind of material, but at the same time he recognizes the possibility of multiple realization, and therefore is not committed to the identity theory.

In arguing that purely formal programs or computational processes are not sufficient for the existence of cognitive states, Searle thinks he has shown, against CTM advocates, that mental processes cannot be purely computational processes. But an even stronger point he raises is that we do not even have good reasons to suppose that computational processes are a *necessary* part of the explanation of intentional mental states. According to him,

> there is so far no reason at all to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally specified elements. As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been given to suppose that they are necessary conditions or even that they make a significant contribution to understanding. (SEARLE, 1980b, p. 418).

In "Is the brain a digital computer?" (1990) and *The rediscovery of the mind* (1992), Searle develops this criticism. He goes on to attack not only strong AI, but also the assumption, accepted by most cognitive scientists, including Fodor, that there is a syntactic level of explanation of mental states and processes (which is a different assumption from the idea that programs are sufficient to explain all aspects of these states). In fact, Searle denies that there is any need to explain the mind by appeal to deeply unconscious mental states and processes. According to him,

there are only two levels of description of the mind: the neurobiological and the psychological, which is necessarily accessible to introspection.

> In our skulls there is just the brain with all of its intricacy, and consciousness with all its color and variety. The brain produces the conscious states that are occurring in you and me right now, and it has the capacity to produce many others which are not now occurring. But that is it. Where the mind is concerned, that is the end of the story. There are brute, blind neurophysiological processes and there is consciousness, but there is nothing else. If we are looking for phenomena that are intrinsically intentional but inaccessible in principle to consciousness, there is nothing there: no rule following, no mental information processing, no unconscious inferences, no mental models, no primal sketches, no 2 1/2-D images, no three-dimensional descriptions, no language of thought, and no universal grammar. (SEARLE, 1992, p. 228-9).

Searle argues, basically, against the explanatory power of the assumption that the brain performs computations. He considers that the traditional notion of computation is such that a computer is simply something to which we can assign zeros and ones, and transitions between those states. Since this is a purely syntactic characterization, and does not refer to any specific material, anything can in theory be a computer, provided it can be assigned syntax. In Searle's view, syntax is not an intrinsic physical property, but is always attributed by an outside observer. And, according to him, one can attribute syntax to anything:

> any object whatever could have syntactical ascriptions made to it. You could describe anything in terms of 0's and l's. (…) the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain. (SEARLE, 1992, p. 208-9).

His idea, then, is that to say that something is a computer is to say something trivial because, in that sense of computing, anything can count as a computer under some description, even a wall. That is, to say that the brain is a computer is not to say something really informative, since one can attribute syntax and transformation of zeros and ones to anything. Moreover, because Searle thinks that syntax is not an intrinsic physical property, but is instead dependent on

an observer, he says that one cannot discover that the brain is a computer, for there is no matter of fact of whether the brain performs or not computations. Since syntax depends on an observer, and is not a physical property, it has no causal power, and plays no real explanatory role in cognition. Computational explanations explain nothing that cannot be explained by neurobiology, which makes them unnecessary to explain mental processes.

3.2 Searle's main points and Fodor's (possible) replies

The paper "Minds, brains and programs" was published in 1980 along with several brief responses against it and, at the end, Searle's replies to these responses. One response is Fodor's, entitled "Searle on what only brains can do."[73] In this section, I will show how Fodor reacts to Searle's central points, and how he could react to Searle's 1990s criticisms of the cognitive sciences.

(1) Programs are not sufficient for intentionality

This is the central point of the Chinese room argument, and Fodor is in perfect agreement with it. In Fodor's words, "Searle is certainly right that instantiating the same program as the brain does is not, in and of itself, sufficient for having those propositional attitudes characteristic of the organism that has the brain. If some people in AI think that it is, they're wrong" (1980b, p. 520). Fodor therefore agrees that the instantiation of a program is not sufficient to give intentional states to a machine, for example, nor to guarantee the existence of intentional states in humans. Programs, or computational processes, do not help us explain the semantic aspect of symbols. In fact, it is interesting to note that Fodor pointed out something similar, even before Searle, in the paper "Methodological Solipsism." According to him,

> People who do machine simulation, in particular, very often advertise themselves as working on the question how thought (or language) is related *to* the world. My present point is that, whatever else they're doing, they certainly aren't doing *that.*

[73] In "Afterthoughts: Yin and Yang in the Chinese room" (1991), Fodor makes a few other remarks against the Chinese room argument. His observations are also followed by Searle's comments. The objection that Fodor raises here is basically that Searle in the room is not a relevant example of instantiation of a program, and therefore cannot be used against the idea that programs are sufficient for intentionality. But I do not intend to discuss this objection here.

> The very assumption that defines their field – viz., that they study mental processes *qua* formal operations on symbols – guarantees that their studies won't answer the question how the symbols so manipulated are semantically interpreted. (FODOR, 1980a, p. 232).

So for Fodor, purely syntactic transformations do not guarantee that the symbols being transformed have meaning, nor do they explain the semantic content of intentional states. Treating mental processes as computational does not tell us anything about the semantics of the states involved in these processes. About that, Fodor and Searle agree.

(2) Mental processes are not processes that operate on purely formal elements

Fodor would also agree with this point, in the sense that, for him, mental processes involve symbols that have not only syntax but also semantics. According to him, "if there are mental representations they must, of course, be interpreted objects" (FODOR, 1980b, p. 520). Mental representations are not purely formal objects, with no semantic content. But it does not follow from that that mental processes should not be treated as, at least in part, *formal operations* on symbols (which are not purely formal). Fodor notes that Searle confuses the two points: "to say that a computer (or a brain) performs formal operations on symbols is not the same thing as saying that it performs operations on formal (in the sense of "uninterpreted") symbols" (FODOR, 1980b, p. 520).

Searle would say that from the point of view of the computer, the symbols it manipulates are not strictly symbols, since they have no semantics for the computer. But it does not follow that mental processes cannot be treated as computational (formal) processes over symbols with semantic content. What CTM says is that mental processes are computational processes, which occur in virtue of the syntactic form, and not the semantics of those symbols. But CTM does not say that mental representations are purely formal, meaningless symbols. The point of RTM (which is compatible with CTM) is precisely that propositional attitudes are relations that we have with mental representations, which clearly have semantic content. CTM does not deny that mental processes are transformations of mental symbols that have semantic content, nor does it say that these processes are intentional because they are computational. CTM is just neutral about the semantics of these symbols.

(3) For there to be intentionality, it is necessary to have something with the causal powers of the brain

Both Searle and Fodor agree that programs are not sufficient for intentionality, and that intentionality is an essential aspect of propositional attitudes. They disagree, however, on what produces it. Searle believes that intentional states are part of our biology, and that they are both caused by the brain, and realized in the brain. According to Searle, from this it follows that only a system with causal powers equivalent to those of the brain could have intentionality:

> I believe that everything we have learned about human and animal biology suggests that what we call "mental" phenomena are as much a part of our biological natural history as any other biological phenomena, as much a part of biology as digestion, lactation, or the secretion of bile. (…) we know that brain processes cause mental phenomena. Mental states are caused by and realized in the structure of the brain. From this it follows that any system that produced mental states would have to have powers equivalent to those of the brain. Such a system might use a different chemistry, but whatever its chemistry it would have to be able to cause what the brain causes. (SEARLE, 1982).

The idea then is that intentionality is a product of the physical structure of a system. In our case, the brain has the power to produce intentionality. Beings with a physicochemical composition different from ours would only have intentionality if their physico-chemical structure could produce it.

Fodor, on the contrary, emphasizes that what is relevant to the existence of intentional states is the existence of appropriate causal relations between the system and the world. He attributes to Searle the idea that the brain is necessary for the existence of intentional states, and that only something with the same physical-chemical structure could give rise to intentionality. As we have seen, Searle avoids being so radical, and recognizes that other types of substrates could produce intentional states.[74] But Searle would say that as far as we know, intentional states are

[74] However, Searle does at times seem to assume, though he tends not to say it explicitly, that only something with the same composition as the brain could give rise to intentionality, and therefore that any attempt to produce intentional states from different materials would be in vain. As we have seen, Searle recognizes that it is possible that other materials can produce intentional states. But he does not really find this possibility very likely (at least not for all materials). Searle says, for example: "I think it is empirically absurd to suppose that we could duplicate the causal

part of our biology, and are caused by the brain. However, in Fodor's view, what is relevant to intentionality is the kind of causal relation that an organism maintains with the world, and not so much its physical composition. Fodor recognizes that our intentional states depend on the causal relations that the brain maintains with the world, but he thinks that from that it does not follow that the physico-chemical composition of the brain is particularly relevant to the existence of these states. As he puts it,

> Searle gives no clue as to why he thinks the biochemistry is important for intentionality and, prima facie, the idea that what counts is how the organism is connected to the world seems far more plausible. After all, it's easy enough to imagine, in a rough and ready sort of way, how the fact that my thought is causally connected to a tree might bear on its being a thought about a tree. But it's hard to imagine how the fact that (to put it crudely) my thought is made out of hydrocarbons could matter, except on the unlikely hypothesis that only hydrocarbons can be causally connected to trees in the way that brains are. (FODOR, 1980b, p. 521).

For Fodor, Searle does not really show why the physical-chemical structure of the brain should be more relevant to giving rise to intentionality than the way we are connected to the world. For that very reason, Fodor accuses Searle of not having a convincing answer to the robot reply. According to him, it is not surprising that Searle, locked in a room, cannot know the meanings of the symbols he manipulates. He has no access to the world, and therefore cannot associate meanings with Chinese symbols. The same is true of today's computers. But that does not mean that a robot, properly connected to the world, could not have intentional states. It just means that machines like Searle in the room, or my laptop, cannot have intentionality because they do not have the means to be in causal relations with the world. According to Fodor, the material of a machine or organism is less likely to be responsible for producing intentionality than the way in which that machine or organism is connected to the environment.

One can imagine a robot that not only instantiates a computer program but also has a specific type of implementation that enables it to have the right kind of causal interaction with the world, whatever it is, in such a way that it has sensors to detect external stimuli, for example, and that it can move around. In that case, would it be so absurd to suppose that this robot could

---

powers of neurons entirely in silicon. But that is an empirical claim on my part. It is not something that we could establish a priori." (SEARLE, 1992, p. 66).

associate meanings to the symbols it manipulates, and that eventually it would have intentionality, even without having a brain? The answer does not seem so obvious, as it is in the case of the Chinese room. We have seen that Searle, in considering the robot's reply, imagines that the inputs of the environment do not give meaning to the manipulated symbols, because they would be just more uninterpreted symbols. In his response to Fodor, he notes that the causal relations between a symbol and the world can only give it meaning if the agent already has intentional states. And, if so, causal relations cannot be constitutive of intentionality, because the agent would already have to have intentionality to be able to interpret a symbol. In Searle's view, intentional states are the product of something that, like the brain, has the power to cause these states, and "outside causal impacts (...) on the formal tokens (...) are not by themselves sufficient to give the tokens any intentional content" (1980a, pp. 522-3). We humans have something with the power of give rise to intentionality, and so we have what Searle calls intrinsic intentionality. Computers, on the other hand, lack a substrate capable of causing intentional states, and therefore the symbols it transforms have intentionality only relative to an external observer. No causal relations between a computer and the world could, in Searle's view, give semantic content to its symbols.

Fodor does not address this last objection from Searle, but we can imagine that he would say that it is not so clear that interaction with the world, together with an appropriate program, could not, as a matter of principle, give intentionality to a system that was not intentional. Searle seems to simply assume that a symbol can only acquire semantic content within a system that is already intentional, but in fact it is not clear that this is true. It is questionable that robots cannot have intrinsic intentional states (not relative to an observer), as long as they have the right implementation, instantiate a suitable program, and have appropriate causal relations with the world. And even if a system already has to be intentional for causal relations to give meaning to a symbol, it does not follow that the causal relations with the world do not play a relevant role in giving meaning to symbols, nor that intentionality is only caused by the physical substrate of a system.

In short, Fodor would agree with Searle that strong AI advocates and some functionalists are wrong in thinking, with Turing, that a computing system, regardless of its implementation and its interaction with the world, may be sufficient for intentionality. A computer that only instantiates a program cannot understand, think, etc., because it only performs formal transformations on symbols, which is not sufficient for intentionality. But Fodor adopts a weaker

thesis, which is that cognitive states may, in theory, arise in machines that implement properly built programs, and that are capable of properly interacting with the environment. It is unclear that Searle has a strong point against this weaker computationalism.

(4) Computational processes are not necessary to explain mental processes

We have seen that in "Minds, brains and programs," Searle says that we have no reason to suppose that computational processes are part of the explanation of mental processes involving intentional states. Although the Chinese room argument is intended only to show that programs are not sufficient for intentionality, Searle speculates that they are not even necessary. Searle thinks, then, that we do not need to assume that there is a level of explanation of the mind that treats mental processes as computational. About this, Fodor accuses Searle of ignoring a vast research program in areas such as linguistics and psychology, which treats mental processes as transformations of symbols. According to him, "to claim that there is no argument that symbol manipulation is necessary for mental processing while systematically ignoring all the evidence that has been alleged in favor of the claim strikes me as an extremely curious strategy on Searle's part" (FODOR, 1980b, p. 521).

In his response to Fodor, Searle ignores this objection. As we have seen, Fodor agrees that computational processes do not give us everything that is essential for mental states. But to do cognitive science, one does not have to accept the idea that computational processes explain all aspects of mental processes, or that the mind is nothing more than a computer program. A more reasonable idea is that we can treat (at least some) mental processes as computational, without assuming that this characterization captures everything that is essential to mental states and processes. And to question this idea, the Chinese room argument is not enough.

We have seen that, in later writings, Searle directly argues against the legitimacy of a computational level of explanation of the mind. Let's see how Fodor could respond to this.

(5) To say that the brain is a computer is to say something trivial. Also, syntax depends on an observer, so it is not an intrinsic physical property, and therefore plays no causal or explanatory role in cognition.

We have seen that Searle thinks that computation, as Turing's conceives it, can be ascribed to anything, so to say that the brain is a computer is to say something trivial. According to him, there is nothing informative in saying that the brain performs computations because, in this sense of computation, anything can be considered a computer. A related problem is that, for Searle, syntax is always assigned from an outside observer, and therefore is not an intrinsic physical property that can be discovered in a system. Basically, we don't need to talk about computations to explain the mind; we only need to talk about neurobiological processes.

Fodor does not address these objections, because they were formulated only in the 1990s, and his comments were directed to the paper "Minds, brains and programs" from 1980. But we can imagine his reply. According to Fodor, computational processes involve syntactic transformations over representations or symbols: "no representation, no computation" (1975, p. 31). Fodor could say, then, that not everything transforms symbols. Maybe zeros and ones can be ascribed to anything, but genuine symbols that stand for something cannot. Crane introduces a useful distinction between instantiating a function and computing a function. According to him, "computing a function, then, requires representations: representations as the input and representations as the output" (CRANE, 2003, p. 103). The same does not occur in the instantiation of a function. He notes that planets, for example, instantiate Newton's laws of motion, but planets "do not 'compute' their orbits from the input they receive: they just move" (CRANE, 2003, p. 103).

Following this distinction, we can perhaps say that not everything does computation. The molecules of a wall possibly instantiate functions, but they do not compute a function, receiving representations as inputs and returning other representations as outputs. We can assign zeros and ones to the wall's molecules, but it is doubtful that these zeros and ones are *representations*, or that a wall computes representations in the same way as a computer or the brain. This narrower conception of computation, which requires representation, could then block the objection that to say that the brain is a computer is to say something trivial, which can be said of anything.[75]

[75] Searle could here demand a more precise characterization of representation. Why is it that molecules in the wall (or thermostats) do not represent anything, but brains do? This is a complicated issue. One thing we could say is that molecules, the wall and thermostats react to things, but that doesn't mean that they do so by employing symbols that stand for something else. As Bermúdez notes, "cognition and information processing come into the picture when there are no direct links between (perceptual) input and (motor) output. Cognitive systems represent the environment. They do not simply react to it. In fact, cognitive systems can react differently to the same environmental stimulus. This is because their actions are determined not just by environmental stimuli, but also by their goals and by their stored representations of the environment. Human agents, for example, sometimes act in a purely reflex manner. But

As Searle himself says, "I do not think that the problem of universal realizability is a serious one. I think it is possible to block the result of universal realizability by tightening up our definition of computation" (1992, p. 209). Searle insists, however, that a narrower definition of computation will not avoid

> *the really deep problem*[, *which*] *is that syntax is essentially an observer-relative notion* (…) We can't, on the one hand, say that anything is a digital computer if we can assign a syntax to it, and then suppose there is a factual question intrinsic to its physical operation whether or not a natural system such as the brain is a digital computer. (…) Computational states are not *discovered within* the physics, they are *assigned to* the physics. (SEARLE, 1992, pp. 209-10).

But if we do accept a narrower notion of computing, not everything will count as a computer, or as something we can assign a syntactic and symbolic level of explanation. And if that is the case, it is not so clear that syntax is entirely dependent on an observer. In my view, the best way to determine whether syntax is an intrinsic property of a system is by considering whether assigning it allows us to better explain the system's behavior. The assumption that a computer follows a certain program clearly seems to serve to explain its behavior. Part of the explanation of why my computer works the way it does involves saying that it instantiates a particular program, and not another. An explanation that was restricted to the hardware, or to the outputs of a computer, would leave out a relevant description of its operation. The same does not happen with an explanation of the behavior of a wall or its molecules. First, because it is questionable that its molecules perform any computation (in the sense of computation as involving syntactic transformations of representations). But even if we adopt a broad enough notion of computation, and accept that we can assign representations to the wall's molecules, and assume that it implements a program, none of this actually serves to explain anything about the wall's behavior. There is no computational theory of walls because nothing is explained by the assumption that the wall's molecules compute anything. In the case of the wall, we have reasons

---

more often we act as a function of our beliefs and desires – not to mention our hopes, fears, dislikes, and so on." (BERMÚDEZ, 2014, p. 282). In order to explain the behavior of molecules in the wall, there is no need to attribute information processing to it. Fodor, in addition, could say that computational processes presuppose a linguistic system of symbols, and that not everything qualifies as a linguistic representational system. A linguistic system is a system whose complex symbols have syntactic structure and compositional semantics. In this way, Fodor could specify that it takes something like a language, a compositional system, to have computation in the relevant sense for cognition. Even if we could assign representations to the wall molecules, it would be questionable that we could assign a compositional representational system to them.

to accept that the attribution of syntax would be entirely dependent on an observer. But in the case of computers, syntax plays a causal role in the system, which indicates that it is independent of any observer.

The question we must ask, then, is whether the assumption that the brain is a computer serves to explain things that we wouldn't otherwise be able to explain. It is at the basis of cognitive science that computations can explain things that pure neurological processes cannot. If this is true then something relevant and informative is being said when we say that the brain is a computer. I think, then, that if the assumption that the brain performs syntactic transformations of symbols allows us to explain the functioning of certain mental processes, as many cognitive scientists accept, perhaps we should say, against Searle, that syntax *is* an intrinsic property of the system, perhaps of a higher order, as Fodor would say. Otherwise, how do we explain the success of these cognitive explanations? There seem to be factual questions about what computations the brain performs, which can be investigated by cognitive scientists, just as there is a matter of fact about which program my computer is now running. In short, if syntax has an explanatory role for cognition, it is probably not entirely dependent on an observer, just as, in some sense, the program that a computer follows is not something that completely depends on an observer.

One of Searle's reasons for saying that syntax is not intrinsic to physics is that "syntax and symbols are not defined in terms of physics. Though symbol tokens are always physical tokens, 'symbol' and 'same symbol' are not defined in terms of physical features. Syntax, in short, is not intrinsic to physics." (1992, p. 225). This leads him to the idea that computations are always attributed from the outside, by an observer, and therefore can never be discovered in the world, nor have causal power. But inferring that something only exists relative to an observer from the fact that is does not have a definition in physical terms seems extremely questionable. Hearts, chairs, airplanes and beliefs certainly don't have definitions in physical terms, and we don't claim that this makes them observer-relative properties. Searle seems to simply dogmatically refuse to accept that syntax can exist and play a causal role in a system because it is not, as it were, there for anyone to see, like the hardware or the outputs of a system. Searle, like Kim, seem to be stipulating that causal powers are completely determined by the entities of lower level sciences, ignoring the fact that special sciences do attribute causal powers to higher-level properties. So in Searle's view, only concrete physical entities can have causal power; symbols don't even exist:

> the difficulty is that the 0's and l's as such have no causal powers because they do not even exist except in the eyes of the beholder. The implemented program has no causal powers other than those of the implementing medium because the program has no real existence, no ontology, beyond that of the implementing medium. Physically speaking, there is no such thing as a separate "program level." (SEARLE, 1992, p. 215).

But Searle misses the point that the level of the program is only one of the levels of description of mental phenomena. Fodor would accept that programs have no causal power as abstractions, but only when implemented. Syntax is, as he says, a higher-level physical property. But still, the fact that syntax does not have a definition in physical terms does not imply its inexistence or causal inertia. As Fodor (1989) notes, we often attribute causal powers to properties that are not definable in physical terms. We say that the wings of a bird, or of an airplane, make them fly, even though wings are multiply realizable, and not definable in physical terms. So "if there's a case for epiphenomenalism in respect of psychological [and, we could add, syntactic] properties, then there is the same case for epiphenomenalism in respect of *all* the nonphysical properties mentioned in theories in the special sciences" (FODOR, 1989, p. 140). I think, then, that Fodor would probably say that Searle has a very narrow view of what might qualify as an entity with causal powers. He would say that if there are computational laws or generalizations, then there are syntactic properties with causal powers, regardless of whether they can be defined in physical terms.

We can, of course, question how well this computational level succeeds in explaining the mind. We may not yet have enough knowledge of the brain and mind to determine whether it is in fact indispensable. But assuming that the computational level is successful in explaining certain aspects of cognition, its explanatory success must be taken, in my view, as an indication of its objective reality.

(6) Computers can't help us understand mental processes because they don't have intrinsic intentionality.

Searle indirectly questions Fodor's idea that computers can help us understand how states with semantic content can have causal powers, since in his view the symbols manipulated by the computer are not really symbols, strictly speaking, because they don't have semantic content.

Computers do not transform symbols that are meaningful to them. Not only the syntax of these symbols, but also their semantics, is dependent on an outside observer. These symbols are only meaningful to the programmers, not to the computers themselves. Thus, we can infer that, according to Searle, we cannot look at the computer to understand how intentional mental processes work, since computers lack intentionality, whereas mental states have intrinsic intentionality.

We have seen that Fodor agrees with the idea that programs are not sufficient for intentionality, and he also accepts that current computers do not manipulate symbols that have meaning to them. He could, however, respond to Searle by saying that programs show us how something that can be understood as a symbol (even if, in the case of the computer, intentionality is attributed from the outside) can have causal powers by virtue of its syntactic form. That the symbols manipulated by the computer have no intrinsic meaning is only a detail, since Fodor does not use the analogy with the computer to explain the intentionality of mental states, but only to explain mental processes, that is, the transition of one mental state to another. In Fodor's view, computational processes can show us how reasoning is mechanically possible; how we can have a causal, mechanical sequence of symbols that is not entirely arbitrary from a semantic point of view. The computer operates through mechanical processes, which are guided by the syntax of the symbols (even if these symbols are only interpreted by programmers). The idea then is that some mental processes could be explained in the same way. The observation that the symbols a computer transforms have no intrinsic intentionality is not enough to challenge this point. Fodor can simply recognize, as he does, that programs are only part of the explanation of mental processes, and that the semantic aspect of intentional states is to be explained in another way, which doesn't involve the analogy with computers (for Fodor, it is by causal relations with the world).[76]

[76] The question remains of how something that is ultimately physical (a network of neurons, for example), can give rise to states that are about something. Even if we accept that computational processes show us how reasoning is mechanically possible, it tells us nothing about how it is possible for something physical to have semantic content. But, in fact, this is a problem for both Fodor and Searle. As Kim notes, "how meaning and understanding could arise out of molecules and cells is as much a mystery as how they could arise out of strings of 0s and 1s" (1996, p. 101).

3.3 Who is right about what?

Having already a general idea of the differences between Fodor and Searle's views, I will now briefly discuss what appear to me to be the correct points highlighted by each of them. I think Searle may be right in emphasizing the importance of the brain for the existence of intentional states. In fact, our only certain examples of systems with mental states are biological organisms with a nervous system (humans and other animals). Everything we know about mental states suggests that they are, at least in part, produced by the brain. From this it obviously does not follow that machines cannot possibly think. But it may be reasonable to assume, with Searle, that their physical constitution will be relevant to ensuring they can have intentional states.

I think, however, that Fodor is right to emphasize that causal relations with the world are relevant to the existence of these states, a point not emphasized by Searle. It seems reasonable to say that our intentional states depend in some way on our interaction with the environment, and that they are not just a product of the brain. But Fodor's view surely is a bit mysterious, for in no place he says what the right causal relations are that give rise to intentionality, nor what sorts of things can be in these relations. And things with brains are all the examples we have so far of things that enter in such causal relations. The best approach, in my view, is to combine the two positions. In our present state of knowledge, the brain, along with how it relates causally to the world, is our best indicator of what is required for there to be intentionality (and also consciousness). It is reasonable to suppose that if intentionality exists in something other than the brain, it will be in part the result of the physical structure of a system, as Searle would say, and in part a product of the kinds of causal connections that the system holds with the world, as Fodor would say. Basically, their views in this respect are not mutually exclusive; they complement one another.

I also agree with Searle in his criticism of the attitude of taking the brain to be irrelevant for the study of the mind. Fodor sometimes exhibits this attitude, perhaps because, as we saw in section 2 of this chapter, he considers that psychology is an autonomous science with respect to neuroscience. But Searle, in trying to emphasize the importance of the brain, and to deny that mental states are realizable in any kind of material, denies the existence of a computational level of explanation of mental states. The two attitudes seem inappropriate to me. As I have pointed out before, it is possible to accept that the brain does in fact contribute in a relevant way to

intentionality, and that studying the brain can help us understand the mind, while accepting that there is an autonomous level of psychological explanation (even if only epistemologically), which makes reference to computational processes and a language of thought.

In the heyday of functionalism, some philosophers and cognitive scientists did think that the computational level captured everything that is essential to mental states, and they tended to say that the reproduction of that level would be enough to produce a system with mental states. This led to the idea, of which Searle is critical, that systems produced from all sorts of materials (valves, a set of doves, or rolls of toilet paper, etc.) could have mental states, as long as they implemented the right program. But it is possible to accept that there is a computational level of explanation of the mind without accepting that it specifies *all* that is essential to mental states and processes, and therefore without accepting unrestricted multiple realizability. We can accept that there is a level of neurobiological explanation, and that there is a level of computational explanation of the mind, without accepting that either one alone captures the complete essence of the mind. Both would be autonomous relative to the other, each with its own laws and entities. There could still be a computational level of explanation of intentional states even if they prove to be dependent on brain states to exist. Moreover, even accepting the epistemological autonomy between the two levels, we can accept that the study of the brain can help us to understand the mind, and vice versa, insofar as our knowledge of correlations between neural and mental states increases. In sum, Searle sometimes seems to want to avoid a computational level of explanation in order to avoid the acceptance of strong multiple realization. But the computational level of explanation does not imply the strong ontological independence of mental states in relation to brain states, i.e. it does not imply that mental states can be realized in materials other than the neurobiological.

Then again, it is possible that Searle is right and intentionality is a product of the brain. Maybe intentional states can only exist in something with the causal powers of the brain. But even if this is true, cognitive science gives us a good indication that computations help us understand mental processes. As Fodor notes, "the cost of not having a Language of Thought is not having a theory of thinking" (1987, p. 147). We will also see in chapter 3 that Fodor argues that the language of thought allows us to explain the phenomena of productivity and systematicity. So, in claiming that there are only two levels of explanation of the mind (common-sense psychology and neuroscience) Searle pays the price of not having an adequate theory of thinking, nor an

explanation of productivity and systematicity. But again, even if Fodor is right and there is a language of thought, whose symbols enter into computational processes, that doesn't mean that mental processes can be implemented in any type of material.[77]

[77] It is worth mentioning that both Fodor and Searle accept that the semantic content of thought is prior to that of natural languages, and that natural language sentences inherit their contents from the thoughts they express (this topic will be discussed in chapter 4). If that is so, it is not unreasonable to say that thought has some kind of structure, prior to the acquisition of language; that it is not an "amorphous mass," as Saussure would say. To suppose that words inherit their semantic content from thoughts is to suppose that there are concepts prior to words, which possibly constitute a language of thought – because they presumably can be combined to form syntactically and semantically complex concepts. Searle, in denying that there is a language of thought, does not seem to adequately explain how natural languages can derive their content from thought.

CHAPTER 3

THE LANGUAGE OF THOUGHT

As I pointed out in the first chapter, one of the initial arguments Fodor presents in favor of the existence of a language of thought is that computational models used to explain cognitive processes presupposed a linguistically structured representational system.[78] At the time of publication of *The language of thought* (1975), this was a rather strong argument, because basically the only cognitive models available treated the mind as a symbol processor. This argument for the language of thought, often referred to as "the only game in town" (e.g. Katz, "The language of thought"), may be taken to lose some of its strength in the 1980s, with the popularization of connectionist models (which are used to explain various cognitive phenomena with no appeal to linguistically structured mental symbols). Although computational theories of mind, which presuppose a linguistic system, are still popular, to the extent that we take connectionist models to be alternative cognitive models, they weaken this argument in favor of a language of thought.

In addition, as I pointed out in the last part of the first chapter, in *The language of thought* Fodor was already pessimistic about the scope of the computational theory of mind, pessimism that is later aggravated in books such as *The mind doesn't work that way* and *LOT 2*. He runs several criticisms against cognitive scientists (such as Pinker) who believe that the mind may be, perhaps entirely, explained computationally. He has argued that CTM can only explain local mental processes, which comprise only a fraction of our mental lives. If this is so, if CTM can only be used to explain a limited number of mental phenomena – such as rational processes involving local properties of mental representations – that are not even the most important and frequent ones (at least not when we consider personal, and not sub-personal processes), it seems that the path from CTM to the language of thought is further weakened. If the only reason to suppose the existence of a language of thought were CTM, and accepting that it only explains some mental phenomena, we would only have reason to believe that the language of thought is involved in some limited mental processes. Presumably we would not need LOT to be involved in the

[78] Another argument was that we need to assume the existence of a language of thought to explain the acquisition of a first natural language. This argument will be discussed in chapter 4.

explanation of inductive reasoning or abductions, for example, if they cannot be addressed in a purely computational or syntactic way.

But there are other reasons, independent of CTM, for assuming the existence of a language of thought. Both Fodor (1987) and Fodor with Zenon Pylyshyn (1988), one of Fodor's main collaborators and also a proponent of the language of thought hypothesis, consider that thought has two properties that need to be explained: *productivity* and *systematicity*. According to them, these properties can only be explained if we assume that there is a system of mental representations which is *compositional*, like natural languages. By language, Fodor basically means a representational system that has constituent syntactic structure and compositional semantics. Roughly, the idea is that thought, like natural languages, exhibits the phenomena of productivity and systematicity, and that a natural way of explaining these phenomena is by the assumption that thought has a syntactic structure and compositional semantics, just like natural languages. Being compositional is what makes thought like a language, hence the *language* of thought. Productivity and systematicity presumably give us reasons to believe that LOT is used at least when we have propositional attitudes, such as beliefs and desires. That is, it would be possible to say that we think in a language of thought even if it is true that CTM is not such a general theory of mental processes.

In *The mind doesn't work that way*, Fodor makes it clear that he believes that these arguments for the language of thought remain strong even with the limitations he attributes to CTM. According to him, "I think that the attempt to explain the productivity and systematicity of mental states by appealing to the compositionality of mental representations has been something like an unmitigated success; in my view, it amply vindicates the postulation of a language of thought." (2001b, p. 04). These are the main arguments Fodor formulates for the language of thought. However, both Fodor and Pylyshyn believe that, even though thought is compositional like natural languages, the language we think in is not the same as any natural language. The arguments for this will be discussed in chapter 4.

In this chapter we shall begin by looking at the two arguments put forward by Fodor and Pylyshyn for the assumption that there is a language of thought (a compositional system of mental representations). In the first part, we will see the argument that goes from the productivity of thought to its compositionality. In the second part, we will see the argument that goes from the systematicity of thought to its compositionality. In the third part, I will indicate how Fodor,

Pylyshyn and McLaughlin criticize connectionist models of cognitive architecture, claiming that they are incapable of explaining the phenomenon of systematicity. If assuming a language of thought allows us to explain phenomena that connectionist models, conceived as cognitive models, don't explain, then they don't pose a strong threat to the LOT. In the fourth part, I will discuss some ideas Fodor presented in a more recent article entitled "Language, thought and compositionality" (2001), in which he goes on to deny that natural languages are compositional. I will argue that the ideas presented in this paper conflict with the arguments of productivity and systematicity, and that they are problematic in their own right.

1. Productivity

The basic idea of productivity is that there does not seem to be a limit to the number of thoughts we can produce and understand. We constantly have thoughts that we never had before, which indicates that our ability to have new thoughts is unlimited.[79] We are obviously finite beings, and therefore the number of thoughts we actually entertain during our lives is finite. But this seems to be a limitation of the finite resources at our disposal, not of our cognitive abilities. Natural languages exhibit the same phenomenon. We constantly produce and understand new sentences, which we had never encountered before, which suggests that there is no limit to the number of possible sentences in a natural language. We in fact only produce and understand a finite number of sentences throughout our lives, but that again seems to be due to limitations in performance, and not in our linguistic competence, to use Chomsky's terminology.

There are, then, two related but distinct questions, regarding the phenomenon of productivity, both in language and in thought. First, what makes it possible for there to be *no limit* to the number of grammatical sentences in a language? In the case of thought, what makes it possible for there to be no limit to the number of thoughts we can have, or, as Fodor puts it, "what is it about beliefs and desires in virtue of which they constitute open-ended families" (1990, p. 17)? How can we arrive, by finite means, at an unlimited number of sentences and thoughts?

[79] In *The language of thought*, Fodor gives a similar argument for the productivity of thought based on the ability we have to consider possible behaviors in new circumstances: "We infer the productivity of natural languages from the speaker/hearer's ability to produce/understand sentences on which he was not specifically trained. Precisely the same argument infers the productivity of the internal representational system from the agent's ability to calculate the behavioral options appropriate to a kind of situation he has never before encountered" (FODOR, 1975, p. 32).

Secondly, how is it possible that we *understand* new sentences, which we have never heard before? Or, in the case of thought, how can we make sense of new thoughts we have never had before?[80] In sum, what we call "productivity" can be divided into two related aspects, the first consisting of the unlimited expressive power of a representational system, the second consisting of our capacity to understand new sentences and thoughts, which we have never encountered before. We can then demand explanations for these two aspects of productivity, both in natural languages and in thought.

As for the first question in language, part of the classic explanation for the unlimited number of possible sentences of a natural language involves assuming that sentences have a combinatorial structure, that is, they are formed from a finite number of constituents (words or morphemes), which are combined according to grammatical rules to form structurally complex expressions. But what really guarantees the lack of a limit to the number of sentences in a given language, as Chomsky would say, is the use of grammatical rules that can be recursively applied. To give just one example, the grammatical rules of English allow sentences to be inserted into sentences repeatedly, as in the sentence "John believes that Mary said that Peter spoke with Sarah about John." In principle, it is always possible to make that sentence longer by adding new constituents, which suggests that there is an infinite number of well-formed sentences in English (and presumably in all other natural languages). Recursive rules, along with the assumption that sentences have a combinatorial syntactic structure, explain the unlimited number of well-formed sentences that can be produced in a language.[81]

We come to the second question regarding the productivity of language: since we can in theory produce new sentences ad infinitum, what guarantees that those sentences can be understood? A common answer, endorsed by Fodor, is that the meaning of a sentence is derived from the meanings of its constituent parts, together with the way they are combined.[82] I can

---

[80] Fodor and Pylyshyn typically characterize productivity only in terms of the absence of the limit of the representational capacity of thought and language. But there are two related phenomena here: that we constantly produce and understand new sentences/thoughts, and that there seems to be no limit to the number of sentences/thoughts we can produce and understand. I will draw on this distinction later in this section.

[81] As Belletti and Rizzi note, "the critical formal contribution of early generative grammar was to show that the regularity and unboundedness of natural language syntax were expressible by precise grammatical models endowed with recursive procedures (…). Over the last fifty years, the technical characterization of the recursive property of natural language syntax has considerably evolved (…). Nevertheless, the fundamental intuition has remained constant: natural languages involve recursive generative functions" (2002, p. 03-04, editors).

[82] This idea has been around at least since Frege ("Compound thoughts"). As Frege puts it: "It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a human being for the very first time can be put into a form of words which will be understood by

understand the sentence "Twenty blue monkeys dance tango in Japan", which I had never encountered before, because I derive its meaning *compositionally*, from the meanings of each of its constituents and the way in which they are combined, which were at least tacitly known by me. Compositionality, as Fodor and Lepore characterize it, "is the property that a system of representation has when (i) it contains both primitive symbols and symbols that are syntactically and semantically complex; and (ii) the latter inherit their syntactic/semantic properties from the former." (FODOR; LEPORE, 2002, p. 01).[83]

The two aspects of productivity are, in fact, two sides of the same coin: the first is essentially a syntactic phenomenon, explained by recursion, whereas the second is essentially semantic, explained by compositionality, but both occur together in language. So sentences having a constituent syntactic structure, recursive rules and compositional semantics is what explains the productivity of natural languages.

Let us now consider the productivity of thought. According to Fodor, "a natural suggestion is that the productivity of thoughts is like the productivity of natural languages, i.e., that there are indefinitely many thoughts to entertain for much the same reason that there are indefinitely many sentences to utter." (1990, p. 18). That is, based on an analogy with natural language, Fodor assumes that the unbounded expressive power of thought "must presumably be achieved by finite means." (FODOR; PYLYSHYN, 1988, p. 33). It seems absurd to assume, for example, that we store an infinite number of mental representations in memory, since our ability to store information is limited.

---

someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought. To be sure, we really talk figuratively when we transfer the relation of whole and part to thoughts; yet the analogy is so ready to hand and so generally valid that we are hardly ever bothered by the hitches which occur from time to time." (1923, p. 01). But whereas Frege takes thoughts to be abstract entities, which are only figuratively structured, Fodor takes thoughts to be psychological entities, which are literally structured.

[83] For discussions on how to better phrase the principle of compositionality, see Szabó (2000, 2010, 2012). In its most general formulation, according to Szabó, the principle of compositionality says that "the meaning of a complex expression is determined by its structure and the meanings of its constituents" (SZABÓ, 2010, p. 255). Another common formulation says that the meaning of a complex expression is a *function* of the meanings of its constituents and the way they are combined. But, as Szabó notes, the formulation in terms of function is too weak: "functions are cheap and determination is not—there is probably a function from the GDP of each country to the number of its bald citizens but the former surely does not determine the latter" (SZABÓ, 2010, p. 256). I will assume that, in saying that a complex expression inherits its semantic and syntactic properties from primitive symbols, Fodor and Lepore simply mean that the semantic and syntactic properties of a complex expression are determined by the semantic and syntactic properties of its constituents. This is similar to the way Fodor states what it is for thought to be compositional: "the content of a thought is entirely determined by its structure together with the content of its constituent concepts" (FODOR, 2008, p. 17).

According to Fodor (and Pylyshyn), the best way to explain the first aspect of the productivity of thought, while taking into account the finite resources of memory, is to suppose that cognition operates through the use of a system of representations which, just like sentences in a natural language, have a syntactic constituent structure and are compositional.[84] So we explain the unlimited number of thoughts we can have by assuming that complex thoughts, just like sentences, are formed from a finite number of simple constituents (concepts), which are combined according to rules, many of which can be recursively applied. But new thoughts are also meaningful. The semantic aspect of productivity is explained by compositionality. Just as the compositionality of language is what explains my ability to understand sentences I have never heard before, it seems reasonable to assume that I can grasp thoughts I have never had before, such as the thought that apple ice cream goes well with coffee, because its meaning is determined by the meanings of its constituent concepts (which are known to me), together with its structure.

Although compositionality is first and foremost a semantic notion, if a representational system is compositional, it will have a combinatorial syntactic structure. This is because the meaning of a complex expression can only be determined by the meanings of its constituents if there are primitive and complex expressions, from a syntactic point of view. Perhaps precisely because compositionality implies the existence of a syntactic structure, in later writings Fodor goes on to simply say that it is compositionality what explains the productivity (and the systematicity, as we shall see in the next section) of thought and language. He says, for example, that "the systematicity and productivity of thought were supposed to trace back to the compositionality of mental representations, which in turn depends on their syntactic constituent structure" (2001b, p. 04). Since compositionality implies a syntactic combinatorial structure, it can be considered part of what explains the productivity of thought and language. However, although compositionality is part of the explanation of productivity, it alone is not sufficient to explain the unlimited number

[84] In Fodor and Pylyshyn's formulation, a system has a combinatorial syntax and semantics when "(a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or are structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure." (1988, p. 12). The item "c" is one way of formulating the principle of compositionality. In that paper, their use of the terms "combinatorial semantics" and "compositionality" is sometimes confusing. Fodor and Pylyshyn argue that what explains the productivity, the systematicity and also what they call there the compositionality of beliefs and sentences is their combinatorial syntax and semantics. By "compositionality" they mean there the principle according to which a lexical item, or a constituent of mental representations, makes "approximately the same semantic contribution to each expression in which it occurs" (FODOR AND PYLYSHYN, 1988, p. 42). Given that this a somewhat idiosyncratic conception of compositionality, I will ignore this step given by the authors.

of representations that a system can produce, because it is possible for a system to be compositional (i.e. to have a combinatorial syntax and semantics), without it having an unlimited expressive power. To explain this aspect of productivity, as we've seen, one needs to say something more about the rules of combination for symbols: that at least some of them can be recursively applied. Fodor tends not to emphasize this point, for he often says that it is compositionality (without reference to recursion) what explains productivity. In short, compositionality alone can explain our ability to understand new thoughts / sentences, but to explain the unlimited expressive power of a system one must also assume that this system has recursive rules.

In any case, the point is that the best way to explain the productivity of thought, according to Fodor, is by the assumption that thought has a very similar structure to that of natural languages. The syntactic constituent structure and the compositional semantics (or, for all purposes, the compositionality, since compositionality entails a syntactic structure) of natural languages is what explains (along with recursion) their productivity. Likewise, the assumption that thought is compositional (i.e., that the meaning of a complex thought is determined by the meanings of its constituent concepts and the way they are combined) would explain the productivity of thought. This means that it is reasonable to suppose that there is a *language* of thought, to the extent that thought, like language, is compositional.[85]

Fodor and Pylyshyn (1988) believe that one possible problem with the productivity argument is that not everyone is willing to accept that our cognitive abilities are unlimited, since we actually only entertain a finite number of thoughts throughout life. They present, then, the argument of systematicity, which avoids the idealization demanded by productivity. It should be noted, however, that though what I have called the first, syntactic, aspect of productivity does involve the assumption that there is an unlimited number of thoughts we can have, the second, semantic aspect of productivity, that is, the fact that we often grasp new thoughts, does not require such an idealization, and is harder to doubt. As Fodor and Pylyshyn do not make the distinction I have proposed between these two aspects of productivity, they do not, in my view,

[85] Note that this argument alone does not show that the language in which thought occurs must be different from any natural language. The argument only shows that the system of mental representations must be compositional (and recursive). Assuming that natural languages are compositional (and recursive), they could in principle be candidates for the language of thought. Fodor formulates several other arguments for the idea that thought does not occur in a natural language. In section 4 and 5, we will see one of his attempts to establish that the content of thought is prior to the content of language, which may be seen as an argument for the view that the language of thought cannot be a natural language. I will also return to this topic in chapter 4.

attribute the proper force to the argument that infers the compositionality of thought from its productivity. Even if one wants to deny that we can grasp an unlimited number of thoughts, the fact is that we often have thoughts we never had before, and it seems natural to ask for an explanation for this phenomenon. We have seen that the assumption that we use a compositional representational system seems to be a good explanation, and that would be so even if the number of thoughts we can have were finite (if we didn't, for example, use recursive rules). So we can preserve the productivity argument for the compositionality of thought even if we want to deny that there are infinite possible thoughts. But let us now look at the systematicity argument.

2. Systematicity

The argument from systematicity is similar to the productivity argument, in that both are based on the comparison of thought with natural languages. The argument, as Fodor formulates it, is that:

(1) Linguistic abilities have the property of systematicity because natural languages are compositional.
(2) Thought is also systematic.
By an inference to the best explanation,
(3) Thought must also be compositional.[86]

Let us first clarify (1). To say that the linguistic abilities to understand and produce sentences are systematic is the same as saying that "the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others" (FODOR; PYLYSHYN, 1988, p. 37). An English speaker who can say/understand, for example, the sentence "Mary loves John", will be able to say/understand the sentence "John loves Mary." Likewise, if someone can say/understand "I think therefore I am", or "the brown cow is taking a

[86] Cf. Fodor, 1987, p. 148. Here Fodor does not use the term "compositional," but says instead that natural languages and thought have a "combinatorial semantics." I take these to be synonyms.

nap", the same person will be able to say/understand "I think", or "the cow is taking a nap".[87] As Fodor and Pylyshyn note, if we learned a language simply by memorizing the sentences we hear, just as when we read and memorize sentences in a phrase book teaching a foreign language, the systematicity of linguistic abilities would be a mystery. In other words, if we assumed that each sentence is an atomic unit, which has no structural connection with other sentences, we could not explain why our linguistic abilities are systematic. There would be no reason to expect that someone who can understand the sentence "John loves Mary" can also understand "Mary loves John". But supposing, on the contrary, that language is compositional, and therefore that sentences have a constituent structure, we can explain why certain sentences are systematically related to others. In the case of the sentences "John loves Mary" and "Mary loves John", both have the same syntactic structure and the same lexical constituents, "John," "loves" and "Mary." We can understand one sentence when we understand the other because we master the grammar of English and know the meanings of the lexical items, which are the same in both cases, and we derive the meaning of each sentence from its syntactic structure and the meanings of its constituents. Likewise, I can produce/understand the sentence "I exist" if I can produce/understand "I think therefore I am" because the former is a constituent of the latter.

As Fodor and Pylyshyn put it, "on the view that the sentences are atomic, the systematicity of linguistic capacities is a mystery; on the view that they have constituent structure, the systematicity of linguistic capacities is what you would predict. So we should prefer the latter view to the former" (1988, p. 38). Thus, Fodor believes that the best way to explain the systematicity of linguistic abilities is by the assumption that not all sentences of natural languages are atomic, but some have a constituent structure, and their meanings are derived in a compositional way from their constituents and the way they are combined.

We come then to step (2) in the argument. The idea is that thinking is also systematic. Fodor and Pylyshyn argue that thought must be at least as systematic as language if we accept (i) that language is systematic, (ii) that understanding a sentence involves entertaining the thought it

[87] Another example, given by Pylyshyn (2003, p. 437), is that if one can say/understand the sentences "Mary hit the ball" and "John baked a cake", then one can also say/understand the sentences "Mary baked the cake" and "John hit the ball" (I'm adapting the example for language, since Pylyshyn gives it for thoughts). It is unclear whether Fodor would accept this. In *Concepts* he says, about the systematicity of thoughts: "the reason that a capacity for *John loves Mary* thoughts does *not* imply a capacity for *Peter loves Mary* thoughts is that they *don't* have the same constituents" (1998a, p. 98). I think, though, that what Fodor is saying is that the capacity to understand (or think) "John loves Mary" does not *alone* imply the capacity to understand "Peter loves Mary". However, if "Peter" is part of my lexicon, and I am capable of understanding "John loves Mary", then I will also be able to understand "Peter loves Mary".

expresses and (iii) that the function of language is to express thought. If understanding a sentence like "John loves Mary" involves having the thought that John loves Mary, and if when I understand this sentence I can also understand the sentence "Mary loves John" (because language is systematic), then if I can have the thought that John loves Mary, I can also have the thought that Mary loves John. If the linguistic ability of understanding sentences is systematic, the ability to have certain thoughts must also be, because the former only occurs because of the latter (since the function of language is to express thought). The systematicity of thought follows from the systematicity of language.

We come finally to point (3) of the argument. Just as the systematicity of language can be explained by its compositionality, so can the systematicity of thought be explained by assuming that we employ a system of mental representations that is compositional and has constituent syntactic structure. To be able to think that John loves Mary makes one capable of thinking that Mary loves John because both mental representations are formed from the same constituents and have the same structure, from which they derive their meanings; the constituents are just combined in different ways.[88] If mental representations were always atomic and did not share constituents and structures, the phenomenon that being able to think that John loves Mary implies being able to think that Mary loves John would be a mystery. It is reasonable to suppose, therefore, that mental representations have a constituent structure and that the meaning of complex representations is derived compositionally from their constituents and the way they are combined, just like sentences of a natural language, and therefore that there is a language of thought.

As for the systematicity of language, one criticism that could be made against Fodor and Pylyshyn's approach is that they give very few examples of systematically related sentences, and no precise definition of systematicity. It is unclear then how two sentences *x* and *y* should be related so that the ability to produce/understand *x* is intrinsically related to the ability to produce/understand *y*. Fodor and Pylyshyn characterize the systematicity of linguistic abilities by stating simply that being able to produce/understand certain sentences implies being able to produce/understand certain others. But given a sentence *x* that I can produce/understand, what

[88] According to Fodor and McLaughlin, "the proposition that the girl loves John is also constituted by these same individuals and relations [as the proposition that John loves the girl]. So, then, assuming that the processes that integrate the mental representations that express propositions have access to their constituents, it follows that anyone who can represent John's loving the girl can also represent the girl's loving John." (1990, p. 94).

are the others that I can also produce/understand? Robert Cummins offers a precise definition of systematicity for language. According to him, "a system is said to exhibit systematicity if, whenever it can process a sentence *s*, it can process systematic variants of *s*, where systematic variation is understood in terms of permuting constituents or (more strongly) substituting constituents of the same grammatical category" (CUMMINS, 1996, p. 23). According to this definition, "John loves Mary" is a systematic variant of "Mary loves John", but also, in the strongest sense, of "Peter despises Cecilia." I think Fodor and Pylyshyn would agree that permutations of constituents, while respecting grammaticality, generate systematically related sentences. It also seems plausible that if I can produce/understand "John loves Mary", *and* I have in my lexicon "Peter", "Cecilia" and "despise", I can also produce/understand "Peter despises Cecilia". To Cummins' characterization of systematic variation, we can add that *x* is a systematic variant of *y* when *x* is a constituent sentence of *y*. That would include cases of sentences like "I think therefore I am" and "I think". If one can produce/understand the former, one can also produce/understand the latter.

Just as Fodor and Pylyshyn do not define the notion of systematicity for sentences, they also do not define it for thoughts. Following Cummins, we could initially say that just as in the case of sentences, two thoughts *x* and *y* are systematically related at least when *x* and *y* have the same constituents permuted. The problem is that in proposing a *definition* of systematicity in terms of constituents, as Cummins notes, it is already being assumed that mental representations have a constituent structure, which was the conclusion to which Fodor and Pylyshyn wanted to arrive. That is, systematicity, so characterized, presupposes that thoughts have constituent structure. However, Fodor and Pylyshyn's argument from systematicity to the constituent structure of mental representations has more precisely the form of an inference to the best explanation. In their view, the phenomenon of systematicity can be explained if we suppose that thoughts have compositional constituent structure, but for them systematicity does not presuppose compositionality. In fact, the same problem appears for the suggested definition of systematicity in language, since the compositional constituent structure seems to be presupposed in the definition.

Perhaps precisely because of this, Fodor and Pylyshyn do not define systematicity in these terms, but only say, as we have seen, that systematicity occurs when the ability to think/produce/understand certain thoughts/sentences is intrinsically connected to the ability to

think/produce/understand certain others, without further specification. Two thoughts are systematically related if the ability to think one implies the ability to think the other. Two sentences are systematically related if the ability to understand one implies the ability to understand the other. However, one might still question that the examples given of systematically related sentences (or thoughts) already presuppose that only those sentences (or thoughts) that share constituents (such as "John loves Mary" and "Mary loves John") are so related. That is, it could be objected that Fodor and Pylyshyn's decision of which sentences are systematically related to a given sentence *x* already presupposes that the sentences are those whose constituents are permutations of the constituents of *x*. But why could it not be said that if I can understand the sentence "Pluto is no longer a planet", I can also understand the sentence "I love chocolate"? After all, I do in fact understand both of these sentences.

What Fodor and Pylyshyn would probably say is that it is possible for someone to be able to understand the first sentence but not the second (and vice versa), but if one can understand "John loves Mary," one can necessarily understand "Mary loves John". As Fodor and McLaughlin note, "cognitive capacities come in clumps. For example, it appears that there are families of semantically related mental states such that, *as a matter of psychological law*, an organism is able to be in one of the states belonging to the family only if it is able to be in many of the others" (1990, p. 92, my emphasis).

It seems perfectly possible for one to be able to think about one's own love for chocolate, while being incapable of thinking about Pluto no longer being a planet. On the other hand, it seems intuitively certain that anyone who can think that John loves Mary can also think that Mary loves John.[89] This phenomenon can then be explained, in the case of the given example, by the supposition that both thoughts have the same syntactic structure and the same constituents, which are combined in different ways. To the extent that they require the same constituents to be

[89] Cummins seems to criticize this idea. According to him, the representational scheme that one chooses to adopt interferes with the intuitions about which thoughts are systematically related: "It is the structure of the mediating representation which determines whether or not we see systematicity in the thoughts." (CUMMINS, 1996, p. 25). What Cummins seems to be saying is that Fodor and Pylyshyn only take the ability to think that John loves Mary to imply the ability to think that Mary loves John because they presuppose that these two thoughts have the same constituents. Possibly, if we adopted another (perhaps non-linguistic) representational scheme, we would not think these thoughts were systematically related. However, it is hard to imagine that, in this case, the intuition that these two thoughts are systematically related would disappear with the use of a non-linguistic representational scheme. Most likely, we would say that the scheme in question does not account for our intuitions - and this is precisely one of Fodor and Pylyshyn's criticisms of connectionism. It is possible that in some cases the representational scheme dictates our intuitions about what thoughts are systematically related, but the intuitive force of the systematic relation between the thoughts that John loves Mary and that Mary loves John seems difficult to deny, and seems to be independent of the representational scheme chosen.

formed, if I can have one of these thoughts, I can have the other. So for Fodor, the idea that systematically related thoughts have permutated constituents would not be part of the definition of systematicity of thought, nor would it be presupposed in the choices of examples, but would be the explanation of this phenomenon. The same would apply to the systematicity of language.

We have seen how the assumption that there is a language of thought is used to explain the productivity and the systematicity of cognitive capacities, and that the explanation of these phenomena in thought is based on an analogy with language. Just as supposing that language is compositional allows us to explain the productivity and systematicity of language, so too supposing that thought is compositional allows us to explain the productivity and systematicity of thought. We do not have direct access to the vehicle of thought, but it is reasonable to suppose that there is a *language* of thought insofar as the assumption of a system of mental representations that is compositional like language would explain these phenomena.

This is not to say that the compositionality of thought depends in any way on the compositionality of language. But both the argument from systematicity and the argument from productivity in favor of a language of thought certainly derive their strength from the assumption that compositionality allows us to explain systematicity and productivity in language.[90] In the way Fodor formulates them, they are deeply grounded in the analogy with natural languages. Fodor and Pylyshyn even say that "beyond any serious doubt, the sentences of English must be compositional to some serious extent" (1988, p. 43-4), and they think there is no "way out of the need to acknowledge the compositionality of natural languages and of mental representations." (1988, p. 45). I'm highlighting these remarks because, as we will see in section 5, Fodor later goes on to deny that natural languages are compositional. In the next section though, I will discuss how Fodor uses systematicity to argue against connectionist models of the mind.

[90] More recently, other arguments have been developed to support the compositionality of language. Pagin (2012), argues that a compositional semantics would help to minimize computational complexity in linguistic communication. So if the semantics of natural languages were compositional, this would help to explain how we usually manage to communicate in an efficient and successful manner. Del Pinal (2015) formulates an argument for the compositionality of the faculty of language based on language acquisition.

3. Connectionism, constituents and systematicity

The central goal of Fodor and Pylyshyn's paper is to defend the classical model of cognitive architecture, as opposed to the connectionist model. According to them, both models accept mental representations. In fact, for them, cognitive models, almost by definition, deal with the *representational* level of explanation of psychological states – and not with the neural level. The difference between the classical model and the connectionist, according to them, is that only the former accepts that mental representations have a compositional constituent structure. Only the classical model accepts that there is a finite number of primitive mental symbols, which can combine to form structurally complex symbols, where the semantic content of complex symbols is determined by the semantic contents of their constituent parts, together with their syntactic structure. That is, only the classical model accepts that there is a language of thought.

Besides, Fodor remarks that "the LOT story amounts to the claims that (1) (some) mental formulas have mental formulas as parts; and (2) the parts are 'transportable': the same parts can appear in *lots* of mental formulas." (FODOR, 1987, p. 137). The classical model therefore accepts that there are mental symbols with a syntactic structure, that the same symbols may appear in different thoughts, and that mental processes are sensitive to the structures of mental representations.[91] Connectionist models according to Fodor and Pylyshyn, only admit non-symbolic representations without a compositional constituent structure.[92] As we have seen in the previous section, the language of thought is taken to be what explains systematicity. Thus, one of the main criticisms that Fodor and Pylyshyn formulate against connectionism is that, because it does not accept representations with constituent structure, it is unable to explain the systematicity of cognitive capacities. Let us see, then, very briefly, what connectionist models say.

Connectionist models of cognitive architecture aim to explain the functioning of the human mind through networks of units that function in a way inspired by neurons. These units are usually divided into layers. The first layer is that of input units, which receive information from outside the system. Next come one or more intermediate layers, which process information

[91] As Fodor and McLaughlin put it, "it is precisely because classical mental representations have classical constituents that they provide domains for structure-sensitive mental processes." (1990, p. 93).

[92] It is worth noting that there is no consensus on whether connectionist models also accept mental representations (although not symbolic). However, as Fodor and Pylyshyn argue, persuasively in my view, if these models do not accept mental representations, it is unclear why they should belong to the cognitive level of explanation, and not to the neural level. I will suggest later that connectionism can also be conceived as an implementation model of classic cognitive architecture.

coming from the input layer and transmit it to units in the output layer, which is the final layer. The units of one layer are connected to units of the next layer, forming large networks of connections.[93] With what is called "training", the force of the connection between units (also called weight) can be adjusted, increasing or decreasing. Each unit has a certain activation value. When this value is reached, a signal is sent to other units in the network, with the intensity being determined by the weight of the connections, forming at the end activation patterns among the different output units.

The idea is then that, in connectionist models, patterns formed by the activation of several units represent something. They form what is called distributed representations. A neural network that is used to classify objects, for example, can use the activation level of input units to represent whether or not a given object has certain features. For example, if the object to be identified is a dog, ideally one or more units will be highly activated if they detect four legs, others if they detect a muzzle, but others will be inactive if they do not detect a beak etc. These units would pass on their activation values and, in the end, the output units would form activation patterns that represent which object is being observed. (With training, the weights of connections among units of different layers can be adjusted so as to reduce the error between a given input and the expected output.)[94]

These distributed representations, however, are not considered to be symbolic. An activation pattern representing a *wet dog*, and another representing a *white dog*, will not usually be composed out of simpler patterns, one of which represents a *dog* and which would be a constituent in both patterns, and which could be transported to take part in other complex representations. They may be patterns that have nothing in common. Patterns of activation are not representations that function as symbols, which can be transported and combined with others to

---

[93] Precisely because they accept networks formed by units inspired by neurons, connectionist models are also called artificial neural networks.

[94] As Waskan explains, "in the simplest case, a particular unit will represent a particular piece of information – for instance, our hypothetical network about animals uses particular units to represent particular features of animals. This is called a *localist encoding* scheme. In other cases an entire collection of activation values is taken to represents something – for instance, an entire input vector of our hypothetical animal classification network might represent the characteristics of a particular animal. This is a *distributed coding* scheme at the whole animal level, but still a local encoding scheme at the feature level. When we turn to hidden-unit representations, however, things are often quite different. Hidden-unit representations of inputs are often distributed without employing localist encoding at the level of individual units. That is, particular hidden units often fail to have any particular input feature that they are exclusively sensitive to. Rather, they participate in different ways in the processing of many different kinds of input. This is called *coarse coding*, and there are ways of coarse coding input and output patterns as well. The fact that connectionist networks excel at forming and processing these highly distributed representations is one of their most distinctive and important features." (WASKAN, "Connectionism").

form complex symbols, while contributing the same meaning wherever they occur. Patterns of activation, or distributed representations, don't have constituent parts, which can be combined with other primitive patterns to form complex patterns, whose meaning is determined by the meanings of the constituent patterns and the way they are combined. Their meanings are therefore not compositionally derivable. According to Sharkey & Sharkey, and Garson:

> connectionist representations are distributed and non-symbolic. As such they are quite different from the symbolic representations employed in the classical approach. (…) Distributed representations combine tokens without those tokens appearing in the complex expression, since the tokens of the input constituents are destroyed in their combination. (SHARKEY; SHARKEY, 2009, pp. 186-187).

> Every distributed representation is a pattern of activity across all the units, so there is no principled way to distinguish between simple and complex representations. To be sure, representations are composed out of the activities of the individual units. But none of these 'atoms' codes for any symbol. The representations are sub-symbolic in the sense that analysis into their components leaves the symbolic level behind. (GARSON, 2015).

In the classical model, on the contrary, complex mental representations are composed out of simpler representations which are always tokened whenever the complex representation is. Fodor and McLaughlin characterize a classic constituent in the following way: "for a pair of expression types *El* and *E2*, the first is a classical constituent of the second only if the first is tokened whenever the second is tokened" (1990, p. 93). In the classic model, the representation *wet dog* will have the representation *dog* as a constituent part, and the same representation may take part in other complex representations. In connectionist models, however, representations always seem to be atomic, in the sense that they don't have constituents that are also meaningful. In addition, the same representation, for example of a dog, can be encoded by different patterns of activation. In the classical model, the same representation will always have the same syntactic form.[95]

[95] As Bermúdez says, "according to the physical symbol system hypothesis [i.e. LOT], representations are distinct and identifiable components in a cognitive system. If we examine a cognitive system from the outside, as it were, it will be possible to identify the representations. This is because physical symbol structures are clearly identifiable objects. If the information a physical symbol carries is complex, then the symbol is itself complex. (…) The information that a distributed network carries is not located in any specific place. Or rather, it is distributed across many specific places" (BERMÚDEZ, 2014, pp. 232-233).

Fodor and Pylyshyn's idea is then that the systematicity of thought is naturally explained by the classical model, that is, by a language of thought, whereas connectionist models don't seem suited to explain it because they do not accept mental representations with a constituent structure. In connectionist models, being able to think of a wet dog does not imply being able to think of a dog. Or, returning to the previous example, the thought that John loves Mary and the thought that Mary loves John need not have anything in common; they can be encoded by entirely different activation patterns. If the mind were connectionist, nothing would guarantee that being able to think that John loves Mary implies being able to think that Mary loves John. But since there are thoughts that are systematically related, Fodor and Pylyshyn say that the mind cannot be connectionist.

Fodor and Pylyshyn's paper has sparked a long debate that remains alive to this day,[96] especially with regard to the issue of systematicity, and it is not my intention to recover it here. It should be noted, however, that there are some who argue that connectionist cognitive models can explain systematicity without accepting structured representations. Cummins et al., in "Systematicity and the cognition of structured domains", recognize that classical models do indeed explain systematicity, at least as far as linguistic understanding is concerned. This is because, according to them, structured representational systems are assumed to have the same structure as the linguistic domain that they intend to represent. But, according to them, it is not necessary to accept representations that are isomorphic to the linguistic domain to explain systematicity. Effects of systematicity on linguistic understanding, for example, require the accepted representational system to preserve and carry information about the structure of the linguistic domain, but this does not mean that the system itself needs to be isomorphic to the linguistic domain. Cummins et al. argue that it could just encode the structure of the desired domain. According to the authors, connectionist models can produce the effects of systematicity in several domains without making use of a structured representational scheme, but using only what they call "structural encodings", which are codifications of the structure of the domain in question, which guarantee that the structure can be recovered. According to Cummins et al.,

> Unlike classical representations that preserve information about the structure of represented elements by actually sharing that structure, tensor-product representations do not share structure with what they represent yet still manage

[96] Cf. Calvo and Symons (2014).

> to preserve structural information and make it available to processors. (CUMMINS et al., 2001, p. 54).
>
> The tensor-product activation patterns themselves do not share structure with their contents, but the existence of recoverability functions for such schemes entails that such schemes preserve structure. Structural encodings permit recovery but do not share structure with their domains. (CUMMINS et al., 2001, p. 62).

The decision of which model to adopt (classical or connectionist) would have to take into account other factors, since both models would be able to explain the systematicity of linguistic understanding: the classical model through a system of representations isomorphic to language, the connectionist through an unstructured representational system, which only encodes information about linguistic structure. What would have structure, in the case of this connectionist model, would be the thing represented, which can be derived from the patterns of activation, but not the representation itself.

The classical model, according to the authors, would have the disadvantage of requiring different representational systems to account for different cognitive domains. The idea is that, since classical representational systems are isomorphic to language, they would only be useful in explaining phenomena involving the linguistic domain. Other cognitive domains (e.g. vision and hearing) have different structures, and therefore cannot be represented by the same system. The connectionist model, on the other hand, admits a system that does not resort to symbolic representations, and that is not isomorphic to a specific domain, so it could be used to encode information from different domains.

On this last point, a possible answer in defense of the classical model would be that it is not clear that representational pluralism is a consequence of the acceptance of the classical model. It is possible that even the systematicity of visual perception, for example, is best explained by a representational system that has a combinatorial structure. It seems that if I can perceive that the cat is on the mat, I can also perceive that the mat is on the cat.[97] That is, it is possible that a

[97] I think this is the kind of systematicity Fodor and Pylyshyn have in mind when they say that, if systematicity were a property only of the cognitive capacities of linguistic creatures, "it would have to be quite usual to find, for example, animals capable of representing the state of affairs *aRb*, but incapable of representing the state of affairs *bRa*. Such animals would be, as it were, *aRb* sighted but *bRa* blind since, presumably, the representational capacities of its mind affect not just what an organism can think, but also what it can perceive. In consequence, such animals would be able to learn to respond selectively to *aRb* situations but quite *un*able to learn to respond selectively to *bRa* situations. (So that, though you could teach the creature to choose the picture with the square larger than the triangle, you couldn't

representational system called linguistic (because it is compositional) operates even in domains that do not involve natural language processing.[98] It sometimes appears that Cummins et al. think that the classical representational system would account only for linguistic understanding. But certainly the use of this system is not so restricted, since the same system would be in operation also when we simply think, when we have beliefs and desires, etc. and not only when, for example, we understand a sentence. It seems to me, therefore, that the scope of the classical model is not as narrow as suggested by Cummins et al.

In response to the central thesis of the authors – that connectionist models also explain systematicity –, it is not clear that the mere encoding of structures can actually account for systematicity. Fodor questions this point in some articles in which he criticizes a connectionist model by Smolensky. According to Fodor and McLaughlin, the constituents of complex mental representations "contribute to determining the causal consequences of tokening that symbol" (1990, p. 105); they can take part in mental processes. In Smolensky's models – which try to account for systematicity without accepting representations with a constituent structure – vectors or complex patterns of network activation (which arise from operations of addition or multiplication of other patterns of activation) are representations without structure, but are decomposable into simpler vectors. As they note, if a complex activation vector encodes a syntactic tree of the sentence "John loves Mary," it does not have in itself the representations of "John," "loves," and "Mary", even if these representations are derivable from the vector representing the sentence. The problem is that the component vectors of a complex vector are not constituents in the classical sense.[99] That is, the components of a complex vector are not there, tokened, when the complex vector is tokened; they are only derivable through operations of decomposition. Although Smolensky's vectors represent structures, they themselves have no structure. And if they do not have structure, they do not have constituents that have causal roles

---

for the life of you teach it to choose the picture with the triangle larger than the square.)" (1988, p. 40-1) "We assume that animal thought is largely systematic: the organism that can perceive (hence learn) that *aRb* can generally perceive (/learn) that *bRa*. But, systematically related thoughts (just like systematically related sentences) are generally semantically related too. It's no surprise that being able to learn that the triangle is above the square implies being able to learn that the square is above the triangle; whereas it would be *very* surprising if being able to learn the square/triangle facts implied being able to learn that quarks are made of gluons or that Washington was the first President of America." (1988, p. 44).

[98] Pylyshyn, in *Seeing and Visualizing*, argues precisely that visualization and imagination, when used in reasoning, require a language of thought.

[99] As Fodor and McLaughlin say, "from the assumptions that bracketing trees have classical constituents and that bracketing trees can be coded by activity vectors, it does not follow that activity vectors have classical constituents. On the contrary, a point about which Smolensky is himself explicit is vital in this regard: the components of a complex vector need not even correspond to patterns of activity over units actually in the machine." (1990, p. 103).

in the system. But, as we have seen, the explanation of systematicity for Fodor and McLaughlin requires that constituents of mental representations can influence the causal sequences of the system. Classical theory accepts that there are mental processes that operate on the constituents of complex representations, for example transporting or recombining them into other complex mental representations. Without constituents being tokened when complex representations are tokened, it is not clear how connectionist models can account for systematicity. According to Fodor and McLaughlin,

> from the point of view of the theory of mental processes, (…) whereas the classical constituents of a complex symbol are, ipso facto, available to contribute to the causal consequences of its tokening – in particular, they are available to provide domains for mental processes – the components of tensor product and superposition vectors can have no causal status as such. What is merely imaginary can't make things happen, to put this point in a nutshell. (FODOR; MCLAUGHLIN, 1990, p. 103).

The point of Fodor and McLaughlin's criticism, then, is that the representational (or, as we shall see, sub-symbolic) level of connectionist networks does not retain in itself the structural relations of the information encoded. Although structured information can be retrieved from patterns of activation through recovery functions, only what is actually exemplified, and not merely derivable, has causal powers. If the structure is not in the representation itself, or in what encodes the information, nothing explains the fact that thinking of a wet dog implies thinking of a dog, or that thinking that John loves Mary implies being able to think that Mary loves John. That is because the connectionist representations of these thoughts don't have, in themselves, tokens of shared constituents, which can be recombined by means of mental processes sensitive to these constituents. [100] Fodor and McLaughlin acknowledge that connectionist models could be

[100] I'm not yet certain of the strength of Fodor and McLaughlin's criticism. If complex vectors are formed by operations that overlap, or multiply, simpler vectors, can't we say that these simpler vectors (or representations) have causal power, insofar as they can combine to form complex representations? Is it really a problem that, after forming a complex vector, the simpler vectors are no longer there to play a causal role? Does the capacity to think that Mary loves John only imply the capacity to think that John loves Mary if the constituents of each of these thoughts are exemplified whenever they occur? Is it not enough for me to be able to think about each of its constituents separately? Perhaps the tokening of constituents is relevant in the case of inferences. The passage from P & Q to Q possibly depends on the process of isolating a constituent of the first thought. Likewise, it seems reasonable to require that the representation of *dog* be tokened when I token a representation of *white dog*, since whenever I have the second thought, I have the first. But for one being able to think *bRa* when one can think *bRa*, it may not be required that the constituents of *aRb* are actually tokened when I have that thought. It may be enough that I have the capacity to think

developed in such a way that, for example, whenever the sentence "John loves Mary" can be processed, so can the sentence "Mary loves John". But if the representations themselves were not structured, this would be a merely accidental feature of the model; there would be nothing in the model itself that would predict systematicity, and therefore explain its nomological character. The classical model, on the other hand, in accepting that mental representations are structured, predicts this type of phenomenon.

Fodor and McLaughlin say then that connectionists must face a dilemma: either they accept a model that admits mental representations without constituent structure, but which doesn't satisfactorily explain systematicity, or they accept mental representations with constituent structure, explaining systematicity, but no longer being an original proposal of cognitive architecture (being at most an implementation model of classic models).

One way to understand connectionism as an implementation of the classical model would be as follows. Given that we can speak of different levels of explanation for cognition, we could say that the classical model is committed to the purely cognitive level of explanation, since it offers a theory of mental representations with no mention of their implementation in the brain. Explanations offered by connectionist models, on the other hand, could be taken as belonging to an intermediate level, in between the purely neurological level and the purely representational or cognitive level. Connectionist networks can then be considered an intermediate step in explaining the physical implementation of mental representations in the brain. But they would not be part of a purely neurological explanation, since they abstract from several physical characteristics of the brain.[101] For example, the units of artificial neural networks have no intrinsic differences between them, whereas there are several types of neurons in the brain, as well as glia cells. We can then conceive of connectionist models as models that intend to explain in a more abstract way how information can be processed by entities such neurons and their networks, while ignoring the details of the biological functioning of the brain. But the explanation of what the representations are, and how they are transformed, would be part of the cognitive level of explanation.

---

a, *R* and *b* in isolation and then put then together in whatever order, even if, once together, they don't wear their structure on their sleeves.

[101] Smolensky, for example, argues precisely that the level of explanation captured by connectionism is what he calls "subsymbolic", in between the symbolic and the neurobiological level: "The subsymbolic level is an attempt to formalize, *at some level of abstraction*, the kind of processing occurring in the nervous system. Many of the details of neural structure and function are absent from the subsymbolic level, and the level of description is higher than the neural level. The precise relationship between the neural and subsymbolic levels is still a fairly wide-open research question; but it seems quite clear that connectionist systems are much closer to neural systems than are symbolic systems." (SMOLENSKY, 1989, pp. 237-8).

The classic model speaks of mental symbols in the hope that eventually these symbols will correspond to something in the brain, but with no intention of showing how these symbols are actually realized in the brain. Connectionist models could then be seen as this intermediate step between the cognitive level and the brain level. A model that used the connectionist apparatus could show the way of this implementation, indicating how symbols could be realized in something physical, such as patterns of activation of units like neurons. It is in this sense, then, that connectionist models could be said to be implementations of the classical model. But this would require showing how *structured* representations can be so implemented.

Contrary to what is commonly thought, the problem that proponents of the classical model have with connectionist representation is not that they are physically distributed. Obviously, Fodor, Pylyshyn and McLaughlin would not argue that each concept is implemented in a specific neuron (like the often criticized idea that there is a specific neuron responsible for representing a grandmother). The problem with the connectionist notion of representation, as we have seen, is that it does not admit structure: each representation is an atom without constituents. The connectionist model, to qualify in the eyes of Fodor, Pylyshyn and McLaughlin as an implementation of the classical model, would have to preserve in its "abstract neurology" structural relations that exist between complex and atomic symbols, for example.

Fodor, Pylyshyn and McLaughlin accept then that connectionism may be an implementation of a cognitive model, but they reject connectionism as an appropriate cognitive model. This is because the way in which connectionism conceives of mental representations is incapable of accounting for the rationality and systematicity of cognitive capacities.[102] Thus, a connectionist model could only be part of the explanation of the systematicity of cognitive abilities if it were at a lower level of explanation than that of the classical cognitive model.

---

[102] Another of Fodor and Pylyshyn's criticism against the connectionist model is that, when it purports to be a cognitive model, it assumes an extreme similarity between the cognitive and the cerebral levels. Some connectionist cognitive models assume, for example, that because representations are anatomically distributed in the brain, at a cognitive level, concepts must be distributed in micro characteristics. But, according to them, there is nothing that forces this kind of isomorphism between levels: "consider, for example, one of the most salient properties of neural systems: they are networks which transmit activation culminating in state changes of some quasi-threshold elements. Surely it is not warranted to conclude that reasoning consists of the spread of excitation among representations, or even among semantic components of representations. (…) The point is that the structure of 'higher levels' of a system are rarely isomorphic, or even similar, to the structure of 'lower levels' of a system. No one expects the theory of protons to look very much like the theory of rocks and rivers, even though, to be sure, it is protons and the like that rocks and rivers are 'implemented in'." (1988, p. 63). So they think it shouldn't be presupposed that a cognitive theory needs to mirror the properties that we discover the brain to have.

Some connectionists think that their models are essentially incompatible with classical models, and therefore that connectionist models could not be implementations of classical models. Some connectionists think, for instance, that classical models require that the rules for manipulation of mental symbols be explicitly represented in the system (in the form of symbols), whereas connectionist models do not accept these rules.[103] This would supposedly show either that connectionist models are different and more plausible cognitive models than the classical ones, or that, even if they were lower-level models, they could not implement a classical cognitive model. In fact, Fodor's computational theory says that mental processes are manipulations or symbol transformations, presumably according to rules. But Fodor denies that the rules themselves need to be explicitly represented. What needs to be explicitly represented are the symbols that occur in mental processes. According to Fodor,

> If the occurrence of a thought is an episode in a mental process, then RTM is committed to the explicit representation of its content. The motto is therefore No Intentional Causation without Explicit Representation. (...) But the rules that determine the course of the transformation of these representations [in mental processes] – modus ponens, 'wh'-movement, 'get the queen out early,' or whatever – need not themselves ever be explicit. They can be emergents out of explicitly represented procedures of implementation, or out of hardware structures, or both. According to RTM, programs – corresponding to the 'laws of thought' – *may* be explicitly represented; but 'data structures' – corresponding to the contents of thoughts – *have to be*. (FODOR, 1987, p. 25).

This point is taken up by Fodor and Pylyshyn. According to them, it is not essential for a classical model that its rules be explicitly represented. The program that specifies the valid transformations of symbol can be an emergent property of the brain, with no need for it to be represented symbolically. According to them,

> one should not confuse the rule-implicit/rule-explicit distinction with the distinction between Classical and Connectionist architecture. (…) What *does* need to be explicit in a Classical machine is not its program but the symbols that it writes on its tapes (or stores in its registers). (…) So, then, you can't attack Classical theories of cognitive architecture by showing that a cognitive process is

[103] According to Sharkey and Sharkey, "connectionism provides an account of the way in which a rule can be inferred without the need for conscious inference, or externally imposed rules or heuristics. Much of the excitement in the psychology community has been about the ability of neural nets to handle apparently rule-governed phenomena without any explicit rules" (2009, pp. 185-6). However, they note that "it is now debated whether such networks still contain rules, albeit implemented in a different manner." (2009, p. 186).

> rule-implicit; Classical architecture *permits* rule-explicit processes but does *not* require them. (FODOR; PYLYSHYN, 1988, pp. 60-1).

Thus, the fact that connectionist models do not represent rules does not show that they are entirely incompatible with classical models or that they cannot be implementations of them – although this may be a disadvantage of the model, in cases where rules are expected to be explicitly represented.

There are also several objections against the classical model which, according to Fodor and Pylyshyn, are irrelevant because they are in fact objections to certain *implementations* of the classical model. These criticisms derive primarily from the common consideration that traditional computers implement classical models. Some connectionists believe, for example, that their model is more compatible with how the brain works, because both the brain and artificial neural networks operate by parallel processing, unlike classic computers that have serial processing. Another recurring criticism is that the brain recovers from damage, a phenomenon that is compatible with connectionist models that accept distributed representations, but not with classical computers. However, none of these observations shows that the classical model is not a good *cognitive* model, and that it cannot be implemented in the brain differently than it is implemented in computers. According to Fodor and Pylyshyn, computers operate serially, but there is nothing that forces classic models to operate in this way. According to them,

> Classical architecture in no way excludes parallel execution of multiple symbolic processes. Indeed, it seems extremely likely that many Classical symbolic processes are going on in parallel in cognition, and that these processes interact with one another (e.g., they may be involved in some sort of symbolic constraint propagation). (FODOR; PYLYSHYN, 1988, p. 55-6).

Nor is there anything to prevent structured symbols from being spatially distributed in the brain. As they see the issue, what is necessary for the implementation of the classical model is that the structural relations between symbols are preserved in physical relations, but the determination of what these physical relations in the brain are, and whether they are distributed or not, is an open question to neuroscience.

Anyone who believes that mental processes involve processing symbols in a language of

thought also believes these symbols are somehow implemented in the brain. Fodor and Pylyshyn recognize that it would be absurd to expect the brain to function in exactly the same way as a digital computer. As I said in the first chapter, Fodor thinks computers can help us understand how intentional states like beliefs and desires can cause other mental states. RTM says that propositional attitudes should be understood as relations with symbols and CTM says that mental processes are computational – and computers show us how symbols can be mechanically manipulated. But that does not mean that Fodor is committed to the idea that the brain has all the characteristics that digital computers normally have. Precisely because of this, the differences in the functioning of the brain and computers cannot be used to attack the classical cognitive model, nor can they be taken as indicative of the incompatibility of connectionist models (as implementation models) with symbolic models. In fact, as Fodor and Pylyshyn say, "the problem with Connectionist models is that all the reasons for thinking that they might be true are reasons for thinking that they couldn't be *psychology*." (1988, p. 66) [104] That is, the reasons given for preferring connectionist over classic models do not seem to be reasons that favor them as cognitive descriptions of the mind. Rather, they favor the idea that connectionist models belong to the level of implementation, and not to the cognitive level.

4. Language of thought and natural languages: some initial remarks

We have so far seen how the language of thought seems to explain the productivity and the systematicity of thought. We may ask what implications the arguments presented in sections 1 and 2 have for the relation between the language of thought and natural languages. That is, we can ask whether the arguments of productivity and systematicity are sufficient to conclude that

[104] It is important to note that it is yet not clear how symbols are implemented and transformed in the brain. Fodor and Pylyshyn do not give us any clue how such implementation could be carried out. They seem to start from the idea that the purely cognitive level, to a certain extent, imposes restrictions on the possible interpretations of the neural level. More than that, at least Fodor sometimes seems to believe that neuroscience can contribute nothing to psychology. We can certainly reject this idea, that the cognitive level of explanation has a complete autonomy with respect to the brain level, so that no discovery of neuroscience can illuminate or impose constraints on psychology. It seems plausible that knowing how the brain works can help us understand how the mind works. But it is important to emphasize that connectionist models, while having the implementation more clearly in their horizon, are still highly idealized. Brain science is still very recent, and in fact it is not yet known how the brain processes information, and how the electrical signals sent by neurons, or the connections formed between them, should be interpreted. Thus, it is questionable that connectionist models, in their current format, can be seen as reliable indicators of how the brain encodes and processes information, although they may be on the right track.

the language of thought must be the same or different from any natural language. The arguments given by Fodor and Pylyshyn show that a compositional representational system allows us to explain the productivity and the systematicity of thought. I believe, though, that neither the idea that thought is productive nor that it is systematic require that the vehicle of thought be a representational system different from that of natural languages. If compositionality were the only requirement for the vehicle of thought, and assuming that natural languages are compositional, natural languages could be candidates for that part. In other words, these arguments do not necessarily show that thought must be different from natural languages, nor that it must be the same; they mainly show that thought must have a compositional structure, something that, so it seems, natural languages have. One might then naturally argue that to explain the productivity and systematicity of human thought it would be simpler to say that we think in a natural language, rather than in a language of thought.

However, it is natural to interpret some of Fodor and Pylyshyn's assumptions in these arguments as already suggesting that LOT must be different from any natural language. In particular, to argue that thought is systematic, they depart not only from the idea that natural languages are systematic, but also from the idea that natural languages have the function of expressing thoughts and that to understand a sentence is to entertain the thought it expresses. Their idea is that the systematicity of language is somehow derived from the systematicity of thought, in that what sentences do is to express thoughts, and that what guarantees the understanding of a sentence is the occurrence of the thought it expresses. So natural languages are systematic because thought is systematic. These two ideas, that the function of language is to express thought and that to understand a sentence is to have the thought it expresses, already seem to presuppose that thoughts are in some sense prior to natural languages, and that the language of thought must be different from any natural language.

Moreover, as we have seen, according to Fodor and Pylyshyn, it seems that the cognitive capacities of non-linguistic creatures are also systematic, so systematicity is not exclusive to the cognitive capacities of linguistic beings. If this is so, then it seems we need a language of thought that is different from natural languages to account for the systematicity of thought in non-linguistic beings. Although their central argument for the systematicity of thought departs from the comparison with the systematicity of natural languages, this comparison is not essential to conclude that thought is systematic. It would be possible to simply say – with not reference to the

systematicity of natural languages – that everything indicates that animals have systematic cognitive abilities and, in the case of linguistic beings, that there are no people who can think, for example, that John loves Mary without being able to think that Mary loves John. As Fodor puts it, "we don't need the systematicity of language to argue for the systematicity of thought" (1988, p. 152).

Fodor and Pylyshyn think that it is useful to compare the systematicity of thought with that of language, possibly because this phenomenon is better understood in language, and because if language is systematic, thought also must be (if natural languages are simply expressing thoughts). But this does not mean that thought is systematic *because* natural languages are systematic, or that it is systematic only in linguistic beings. On the contrary, the dependence is in the opposite direction: the systematicity of language depends on the systematicity of thought.[105] My point, then, is that although one might initially think that the explanation of the systematicity of thought by compositionality allows one to say that natural languages may be the vehicle of thought, the natural interpretation of the systematicity argument, as formulated by Fodor and Pylyshyn, is that thought is different from any natural language. This is because they assume that the function of natural languages is to express thoughts, that to understand a sentence is to have the thought that it expresses, and that the perceptual and cognitive capacities of animals are also systematic. These points suggest that the vehicle of thought must be a language different from any natural language.

However, it is also possible to read the systematicity argument as not presupposing a difference between LOT and natural languages. That is, one might deny that the assumptions that language expresses thought and that understanding a sentence is to entertain the thought it expresses lead to the conclusion that the vehicle of thought is not a natural language. One could accept (1) that the vehicle of thought is a natural language; (2) that natural languages, besides serving as the vehicle of thought, also have the function of expressing thought and (3) that the thought that constitutes the understanding of a sentence involves the sentence. So the assumption that the function of language is to express thought alone would not imply the thesis that the

[105] Searle uses a similar strategy in his treatment of intentionality in thought and language. He begins his treatment of intentionality by its manifestation in language, but he claims this is a heuristic procedure. In what he calls "direction of logical analysis," the intentionality of mental states comes first: "In my effort to explain Intentionality in terms of language I am using our prior knowledge of language as a heuristic device for explanatory purposes. (…) Language is derived from Intentionality and not conversely. The direction of pedagogy is to explain Intentionality in terms of language; the direction of logical analysis is to explain language in terms of Intentionality." (SEARLE, 1983, p. 05).

vehicle of thought is not a natural language. According to this view, natural languages could have two functions: being the vehicle of thought and the expression of thought.[106] Even though Fodor and Pylyshyn are committed to thought being different from any natural language, one could still accept productivity and systematicity arguments while denying this view.

Therefore, in order to establish the difference between LOT and natural languages, we need other arguments. In the next chapter, we will see stronger arguments put forward by Fodor, Pylyshyn, and Pinker for the idea that the language of thought is not a natural language. In the following section, I will critically examine a more recent article by Fodor, in which he formulates another argument for the idea that the content of thought is prior to the content of natural languages. I will also argue that the ideas present in this paper conflict with the productivity and systematicity arguments formulated in the opening sections of this chapter, and that they are problematic in their own right.

5. Thought, language and compositionality

We have seen in the first two sections of this chapter that the language of thought is supposed to explain the productivity and systematicity of thought, and that the explanation of these properties in thought is based on the explanation of these properties in language. As we have seen, the idea is that language is productive because we can in theory produce and understand an infinity of new sentences, and it is systematic because the ability to produce and understand certain sentences is directly related to the ability to produce and understand certain others. These phenomena are then explained by the assumption that language is compositional. Fodor and Pylyshyn argue that just as productivity and systematicity are explained in natural languages by compositionality, analogously, the assumption that the system of mental representations is compositional seems to offer the best explanation for the productivity and systematicity of thought. It is natural to infer that thought is compositional because this assumption seems to work well to explain the phenomena of productivity and systematicity of language.

In light of the productivity and systematicity arguments, it is somewhat surprising that in

[106] However, it is unclear what the proponent of this view would say about the systematicity of cognitive capacities of animals.

the article "Language, thought and compositionality" (2001) Fodor goes on to argue that natural languages are not compositional. In this section, I will present the central ideas of this article, and then raise three worries about it.

5.1 Main ideas

Do sentences mean what they do because of the thoughts they express, or is it thought that derives its semantic content from the words and sentences of a natural language? Fodor (2001) has claimed to have a solution to the traditional problem of what comes first (in order of explanation of semantic content), thought or language. We can call it the thought-or-language problem. Fodor assumes that either sentences derive their content from thoughts, or vice versa, without a third option. His proposal, basically, is that the issue should be decided by compositionality. According to Fodor, any adequate theory of content must accept that content is compositional: "one thing that we know about content *for sure*: It is compositional in whatever it is underived in" (2001a, p. 06). Compositionality is, in Fodor's words, non-negotiable. Fodor thinks this for the reasons we saw in the previous sections: it is compositionality that explains productivity and systematicity. So whatever has content in the first place, thought or language, must be compositional. Fodor characterizes compositionality here by saying that "the semantic value of a thought (/sentence) is inherited from the semantic values of its constituents, together with their arrangement" (2001a, p. 06). He also notes that a complex thought or sentence will only be compositional if it is explicit about its content and structure, that is, if its content contains *all and only* the semantic properties that it inherits from its constituents and structure. So in his view, compositionality requires a straightforward correspondence between the constituents of a thought/sentence and the constituents of its content. Given that compositionality is non-negotiable, if only one of thought and language is compositional, this will be the one that has semantic content in the first place.

Fodor then argues that language is not compositional. He starts from the premises that sentences express thoughts, and that the content of a sentence is the content of the thought it expresses. According to Fodor, if language were compositional, sentences would have exactly the same constituent structure as the constituent structure of the thoughts they express. That is, for language to be compositional, it is necessary, in Fodor's view, that sentences be explicit about the structures and contents of the thoughts they express; there must be a straightforward

correspondence between the constituents of a sentence, and the constituents of its content, that is, of the thought it expresses. If language were compositional, the meaning of a sentence (i.e. the thought it expresses) would be derived only from the meanings of its constituents and the way they are combined. However, according to Fodor, it is an empirical fact that sentences are often inexplicit and elliptical about the structure and content of the thoughts they express.

Fodor gives as an example the sentence "it's three o'clock," when uttered as an answer to the question of what time it is. Imagining that the sentence is uttered at three o'clock in the afternoon, the thought that one intends to communicate is that it is three o'clock in the afternoon here and now, but the sentence is inexplicit about the time of day and the place. If we were to derive the meaning of the sentence in a purely compositional way, from its constituent parts, we would not arrive at exactly the same thought that it is normally used to express, namely the thought that it is three o'clock in the afternoon here and now. No constituent of the sentence "it's three o'clock" corresponds to *afternoon*, for instance. This is why the same sentence can be used to express both the thought that it is three o'clock in the afternoon and the thought that it is three o'clock at night. So the structure of the sentence leaves out something that appears in its content. Since compositionality requires explicitness, the sentence is taken to be non-compositional.

Fodor's point is that we often use sentences that are not entirely explicit about their contents, that is, about the thoughts they are used to express. It seems then that in many cases the content of a sentence is not determined compositionally, by the meanings of its constituents and the way they are combined, because sentences often do not have all the constituents of the thoughts they express. If language were compositional, this would not be the case. According to Fodor, "either the content of the thought is different from the content of the sentence that expresses it, or the sentence isn't compositional. I take it that the first disjunct is preposterous; so I take it that the second disjunct must be true" (2001a, p. 12).

This argument alone does not show that thought is compositional. But if we accept that either language or thought must have content in the first place, that compositionality is what decides which content comes first, and that language is not compositional, we are led to the idea that thought must be compositional. To support this idea, Fodor says that the problems raised against the compositionality of language do not apply to thought. According to him, "whereas the content of a sentence may be inexplicit with respect to the content of the thought it expresses, a thought can't be inexplicit with respect to its own content; there can't be more—or less—to a

thought than there is to its content because a thought just *is* its content" (2001a, p. 14). The idea seems to be that sentences can be inexplicit because they, strictly speaking, have no content of their own but derive their content from the thoughts they are used to express. This creates the possibility that there may not be a direct correspondence between the constituents of a sentence and the constituents of its content. The same inexplicitness cannot happen with thought, because, as Fodor puts it, a thought just is its content. Fodor is assuming that a thought does not derive its content from something external to it, and so it is not subject to being inexplicit about its content; it is not possible for a thought to have more or fewer parts than there are in its content. Fodor seems to conclude from the explicitness of thought that thought is compositional. According to him, "a mental representation is ipso facto compositional with respect to the content that a correct semantics would assign to it" (2001a, p. 14). Once it is accepted that thought, unlike language, is compositional, Fodor concludes that the content of thought, unlike the content of language, is not derived, or is the one that comes first in order of explanation.

5.2 Problems

The arguments and ideas presented in "Language, thought and compositionality" can be questioned in several ways. I will formulate three problems here. The *first* is that the idea that language is not compositional affects the arguments formulated earlier by Fodor in favor of the existence of a language of thought. As we have seen, Fodor said that we can explain the productivity and systematicity of thought by the assumption that thought is compositional, just as we explain the productivity and the systematicity of language by assuming that language is compositional. These arguments were entirely based on the analogy with language. But if the inference from the productivity and systematicity of language to its compositionality no longer works, since Fodor now says that language is not compositional, what guarantees that it works for thought? If natural languages are not compositional, their compositionality cannot be what explains their systematicity and productivity. Similarly, it seems that compositionality could not be what explains the systematicity and productivity of thought.

It is certainly possible that thought is compositional and language is not. Fodor never intended to say that the compositionality of thought depends on the compositionality of language, for example, or that if language is not compositional, then thought cannot be either. But still, if language is not compositional, Fodor owes us a reformulation of the productivity and

systematicity arguments for the language of thought, which does not rest on the analogy with natural language.[107] Also, what led to the conclusion that the system of mental representation was language-like was that it was compositional, like natural languages. If thought is no longer considered compositional by means of the analogy with language, since language is no longer compositional, why continue to say that there is a *language* of thought?

As we saw in section 5.1, it is possible to read Fodor as sketching an argument for the compositionality of thought in that same article. The idea is that thought is compositional because it is always explicit about its content. So maybe this could be a way of arguing for the existence of a compositional system of mental representation that would not depend on the analogy with language. A first problem here is that Fodor assumes that thought is always explicit (that there isn't anything extra or missing in a thought with respect to its content) because he is already assuming that thought doesn't derive its content from something else. The problem with this assumption is that the argumentation is circular regarding the thought-or-language problem, as I will argue later.

But in addition, at this point it seems that Fodor is assuming that, because it is explicit, thought must be compositional. This is a questionable step. Even accepting that thoughts are always explicit (that there isn't more or less to a thought than there is to its content), and that thoughts do not derive their content from something external, it doesn't follow from this that thoughts are composed of parts (concepts), which combine to form complex formulas in a language of thought, which derive their content from their component concepts and the way they are combined. There is nothing that seems to exclude *a priori* the possibility that all thoughts are atomic, not compositional, units, which are always explicit about their contents. Explicitness only requires that there is no misalignment between the structure of a thought and the structure of its content, but it says nothing about complex thoughts being constructed out of simple units. That is

[107] As I will mention below, Fodor later (2008) goes in the direction of denying that natural languages are, strictly speaking, productive and systematic. That could be a way, even if not a very satisfying one, of preserving the inference from productivity and systematicity to compositionality for thought, while denying that the compositionality of language is what explains its productivity and systematicity (for, strictly speaking, language is neither productive, nor systematic). Even though Fodor's new view, that language is not compositional, conflicts with the productivity and systematicity arguments for the language of thought, I'm leaving it open that the arguments could be preserved after some reformulation such as the one just mentioned. For a stronger opposition against Fodor's arguments for the compositionality of thought, see Clapp (2012). Clapp also notes the problem that denying language's compositionality raises to the systematicity and productivity arguments for thought, but he goes on to argue that neither language nor thought is (truth-conditionally) systematic nor (truth-conditionally) compositional. While Clapp may be right, as I read Fodor, he is committed only to meaning, and not to truth-conditional, systematicity and compositionality.

an extra assumption which, though reasonable, cannot be inferred from thought's supposed explicitness. So even if the compositionality of thought requires explicitness,[108] compositionality does not follow from the assumption that thought is always explicit, unless it is already taken as a starting point that for a thought to be explicit is for it to have a constituent structure that determines a thought's content (in which case, the compositionality of thought would have been simply assumed together with its explicitness, and not argued for).

To be sure, unlike Clapp (2012), I'm not trying to argue that thought is not compositional. According to Clapp, by accepting that language is not compositional, Fodor must also accept that thought is not compositional. The problem is that Clapp's notion of compositionality rests on the assumption that the semantic content of declarative sentences and thoughts is their truth-conditions, or the propositions they express, to which Fodor is not here explicitly committed. So in arguing against Fodor's explicitness argument for the compositionality of thought, for instance, Clapp assumes that a thought's content is a proposition, which is conceived of as a set of conditions under which a thought is true. He then goes on to argue that there are cases where it seems unlikely that a thought is explicit about all that appears in the proposition it expresses, which means that the proposition a thought expresses will not always be derived compositionally. But Fodor is in that paper assuming that there isn't a separation between a thought and its content. He is not assuming that thoughts or sentences derive their content from a third thing, namely a proposition, but rather that the semantic content of a sentence is the thought it expresses, and that a thought just is its semantic content (and not its truth-conditional content, as Clapp claims). However problematic Fodor's views might be (it is certainly odd to say that a thought just is its content), they are not, I think, subject to Clapp's criticism, for they don't share the same assumptions. In addition, there seems to be an intuitive appeal to the idea that a thought's meaning doesn't need to be identical to a proposition that specifies its truth-conditions. So even if one accepts Clapp's argument that thought's truth-conditional content is not compositional, there could still be another form of content that is, as Clapp himself admits (2012, p. 321).

All I'm claiming, then, is that in accepting that language is not compositional, we are left with no real argument for the compositionality of thought, and therefore for the idea that there is a language of thought. The explicitness argument can be questioned first because Fodor gives no

[108] For a criticism of the idea that compositionality requires explicitness, see Elugardo (2005).

argument for the explicitness of thought, and second because he seems to be simply assuming that compositionality follows from explicitness. The second assumption, as I've tried to show, is false. But that doesn't mean, as Clapp suggests, that we should reject the compositionality of thought. We could either formulate new arguments for it, or we could try to preserve the systematicity and productivity arguments by reestablishing the compositionality of language. In the final section, I'll indicate how the second alternative could be pursued.

It could be said, though, that Fodor is arguing for the compositionality of thought not solely on the basis of the explicitness of thought, but rather on the basis of thought being productive, systematic *and* explicit.[109] So compositionality can be inferred for thought, even if it can't for language, because thought, besides being productive and systematic is, unlike language, also explicit. But, assuming that productivity and systematicity are not enough to infer compositionality (since language is supposed to have both of these characteristics, while still not being compositional), it is unclear why adding explicitness to the picture would make the case for the compositionality of thought any stronger. It seems that the most the assumption of explicitness could do is to create no impediments for the idea that thought is compositional (since the fact that language is not explicit is supposed to be a problem for the assumption that language is compositional). But adding that thought is explicit seems to leave both the argument from productivity and the argument from systematicity unaffected. And, as I suggested earlier, there is no need to assume that thought being explicit makes a case for thought being compositional.

Besides, it is important to stress that we were not initially looking for an argument for the compositionality of thought or language for its own sake, but rather we were looking for an explanation of productivity and systematicity. Compositionality was supposed to be what explains productivity and systematicity, both in thought and in language. Even if we concede that the explicitness of thought makes a stronger case for the compositionality of thought, if compositionality is not what explains the productivity and systematicity of language, it is unclear how it could be what explains these features of thought. And assuming that we accept that thought is compositional and that compositionality is what explains the productivity and systematicity of thought, we would still be left with the problem of explaining the productivity and systematicity of language. Assuming that language is not compositional would leave these characteristics unexplained.

---

[109] I owe this observation to an anonymous reviewer from *Philosophical papers*.

This is the *second* problem with Fodor's new views. Fodor often argues, including in this very paper, that compositionality explains the systematicity and productivity of both thought and language.[110] Denying that language is compositional makes its productivity and systematicity a mystery. What Fodor could do is deny that natural languages are, strictly speaking, productive and systematic. In *LOT 2* (2008) Fodor seems to hold precisely that view: the productivity and the systematicity of language are only apparent; they are parasitic on the productivity and the systematicity of thought. According to him, "one can imagine a view according to which *only* thought is compositional in the first instance and the apparent productivity, systematicity, etc. of languages is parasitic on that of the thoughts they are used to express. In fact, I'm inclined to think that's the right view." (2008, p. 55, n. 8). What Fodor seems to mean by this is that language is only productive and systematic because thought is productive and systematic.

In the case of the productivity of language, the idea would be that there is no limit to the number of grammatical sentences of a language because there is no limit to the number of thoughts one can have. That is, language is only productive because thought is productive. But even if this is true, that is, even if the productivity of language derives from, or is parasitic on, the productivity of thought, it still must be explained. And it is not clear how the productivity of language could be explained only by the compositionality of thought. It seems that our capacity to understand new sentences, and produce an unlimited number of grammatical sentences of a language must be, at least in part, a consequence of the mechanisms of language itself. And the natural assumption is that language itself is, to some degree at least, compositional.

Likewise, even if the systematicity of language is derived from the systematicity of thought, it seems reasonable to suppose that the English sentences "John loves Mary" and "Mary loves John" are systematically related because their meanings are obtained in a compositional way, from words and structures that they both share. It is not clear how the relationship between these two sentences could be explained by appealing only to the compositionality of their corresponding thoughts, and not to the compositionality of the sentences themselves. That is, it seems strange to suppose that they are systematically related only because their mental correlates

[110] Strangely, perhaps by a relapse, in arguing for the non-negotiability of compositionality, Fodor says in this same article that "both human thought and human language are, invariably, productive and systematic; and the only way that they could be is by being compositional." (2001a, p. 06). He also says that English being compositional is what explains the possibility of forming several different sentences about, for example, doves and the weather in Manhattan, with the same words being used with the same meanings in different sentences. Apparently, Fodor changes his mind a few pages later, when he goes on to say that "as a matter of empirical fact, language is pretty clearly not compositional." (2001a, p. 11).

are, and that there is nothing in language itself that guarantees the systematic relationship between these two sentences. In short, if Fodor wishes to defend the idea that language is not compositional, he must offer some other explanation for the productivity and the systematicity of language. Even accepting that these phenomena are parasitic on the same phenomena in thought, it seems reasonable that they are to be explained by the compositionality of language, even if it is also dependent on, or parasitic on, the compositionality of thought.[111]

The *third*, and most substantial problem with Fodor's views is that his central argument is circular.[112] As we have seen, Fodor's goal is to determine what comes first in order of explanation, the content of thought, or the content of language. According to him, this question must be decided by compositionality. Whichever is compositional will be the one that has underived content. However, in arguing that language is not compositional, Fodor assumes that the function of language is to express thought, and that the content of a sentence is the content of the thought it expresses. As we have seen, he says that sentences often have more or fewer constituents than the thoughts they express. Because sentences are often inexplicit about their contents, they will not be compositional. But in this argument, Fodor is already assuming that sentences have no content of their own, and that the content of a sentence is not independent of the content of thought. That is, in arguing against the compositionality of language, he already supposes that the content of a sentence is to be explained in terms of the content of the thought it expresses. And if this is so, we already know that language cannot come first in the explanatory order, even before

---

[111] Szabó (2010) argues against the arguments that infer the compositionality of linguistic content from the productivity and systematicity of language. He adopts a Kaplanian distinction between character (understood as linguistic meaning out of context) and content (linguistic meaning relative to context). According to him, productivity and systematicity only give us grounds to accept the compositionality of character, but not the compositionality of expression content. Even if he is right, some compositionality (namely, of meaning out of context) still needs to be attributed to language in order for its productivity and systematicity to be explained. But it is not clear that Szabó is right about productivity and systematicity not giving us grounds to infer the compositionality of expression content (assuming we accept that distinction). His point against the productivity argument, for instance, is that "it is simply not true that competent speakers can in general understand—know the content of—complex expressions they never encountered before purely on the basis of their linguistic competence." (SZABÓ, 2010, p. 261). He highlights the fact that we typically understand new sentences in contexts of utterance, and that knowing the character of the constituents of a new complex expression is not sufficient to determine the expression's content relative to a context. But the proponent of the productivity argument could reply that what explains our ability to understand the content of a new complex expression in a context is the compositionality of expression content: the content of a complex expression is determined by its structure and the *contents* of its constituents, which in turn can be sensitive to the context. The view is not, as Szabó characterizes it, that the *character* of the constituents of a complex expression is sufficient to determine the expression's content, but rather that the *content* of the constituents (which can depend on the context) is.

[112] Other discussions of Fodor's argument have focused on whether language is in fact compositional (ELUGARDO, 2005; SZABÓ, 2010), and whether thought itself is compositional (CLAPP, 2012), but they have failed to notice what I take to be the central flaw in Fodor's argument, namely, its circular solution to the thought-or-language problem.

we know whether language is compositional or not.

The same circularity occurs when Fodor argues that thought, unlike language, is compositional. The idea is that thought cannot be inexplicit about its content, because the content of a thought is the thought itself. Given that all that is in a thought is in its content, since they are the same thing, thought is always explicit and, presumably, compositional. But this argument presupposes that the content of thought does not derive from something external to it, for example the content of the sentences of a language. And if this is so, we already know in advance that it is thought that has underived content, and not language, even before we know whether or not thought is compositional.

In the way that Fodor argues, the explicitness and the compositionality of thought are in fact consequences of the *assumption* that thought has underived content. Fodor comes to the idea that thought is compositional because he has already stipulated that the content of thought is explanatorily prior to the content of sentences in a natural language. If we withdraw this stipulation, neither his idea that thought is explicit nor that it is compositional holds. But Fodor could only use the compositionality of thought to determine in a neutral way that thought has content in the first place if his argument in favor of the compositionality of thought were independent of this presupposition, and therefore not circular. Perhaps explicitness and compositionality do follow from the assumption that the content of thought is underived. But we would then need an argument in favor of the underived content of thought other than the argument from compositionality, one that does not presuppose underived content, and Fodor does not offer us one. I'm not saying there aren't any good reasons to think that language derives its content from thought. There may be, but compositionality is not one of them. Fodor cannot legitimately use compositionality to decide whether it is thought or language that has underived content because he is using the idea that language has derived content to argue against its compositionality, and the idea that thought has underived content to argue in favor of its compositionality.

5.3 Possible replies

One possible way out of the first two problems presented above (namely, that we are left with no arguments for the compositionality of thought and with no explanation for the productivity and systematicity of language) is to try to reestablish the compositionality of

language. That way, we could preserve the productivity and systematicity arguments for the language of thought, as well as explain the productivity and systematicity of language. There are several ways to do this.

We have seen that Fodor's main point against the compositionality of language is that sentences are often inexplicit about their contents. But one could deny that elliptical sentences such as "it's three o'clock" are really inexplicit about their contents. That is a common strategy adopted, for instance, by Elugardo (2005). As he notes, most theories about syntactic ellipses adopt the distinction between the surface form and the logical or deep syntactic form of sentences, and hold that only the first is inexplicit. According to one view, the sentence "it's three o'clock" would also have "here" and "now" as hidden syntactic constituents, even though its surface form does not make these constituents explicit. If we accept that this sentence has a hidden syntactic form that is explicit about its content, Fodor cannot use the inexplicit character of its superficial form to argue against the sentence's compositionality.

But against that, I suspect Fodor would say that assuming that the sentence "it's three o'clock" has hidden constituents that are not, as it were, there for everyone to see, is a somewhat ad hoc stipulation, whose only motivation is to preserve the sentence's compositionality.[113] Instead, if there really are hidden constituents somewhere, in Fodor's view they are constituents of the thought that is being expressed. As he says in *LOT 2*, "I think that LF [logical form] is a level of description not of English, but of Mentalese" (2008, p. 78, n. 50). Assuming that sentences are used to express thoughts, there is no reason to suppose that there will always be a correspondence between the constituents of a sentence and the constituents of the thought that it is expressing. In order for communication to occur efficiently, we may say less than what we think, because, e.g., information that is of mutual knowledge between speaker and hearer doesn't need to be made explicit in a sentence, when it is not relevant. In some of these cases at least, Fodor could insist that the deep structure of the sentence is better understood not on the level of the sentence, but rather as a specification of the constituents and structure of the thought being expressed, which the sentence doesn't encode completely due to conversational maxims that speakers and hearers implicitly follow. So, in the case of "it's three o'clock", claiming that its deep structure is

[113] Fodor says that "the more or less patent uncompositionality of English isn't something that formal semanticists like to admit. Rather, as far as I can tell, it's the house rule in formal semantics that you hold onto the compositionality of natural language, whatever the cost in face implausibility may be." (FODOR, 2001, p. 13). Szabó seems to share this concern, when he points out that "the fixes semanticists come up with when faced with putative counterexamples to compositionality are often complicated and lack independent motivation" (SZABÓ, 2012, p. 73).

compositional would be the same as claiming that the thought expressed by it is compositional. The sentence proper, to the extent that it lacks constituents which are present in the thought that it expresses, would not be compositional according to Fodor.[114]

Whether or not Fodor is right about this, there are still other ways of challenging his claim that language is not compositional.[115] Another possible way out, while still accepting the idea that the meaning of a sentence is solely the thought it expresses, would be to admit that there are degrees of compositionality in language. While sometimes there may be more in the meaning of a sentence than what can strictly be extracted from its constituent parts, as in the case of the sentence "it's three o'clock" (assuming we reject the postulation of hidden constituents in this case), Fodor hasn't shown that this is usually the case. Other sentences could be explicit and compositional. And even if it is true that sentences are not, for the most part, entirely explicit about their contents, it seems at least sometimes possible to transform inexplicit sentences into sentences that are explicit about their contents. One can say, for example, that "it is right now three o'clock in the afternoon here in Paris". Of course, one doesn't usually say this because in communication we typically follow certain maxims, such as the conversational maxim of quantity stated by Grice (1975), that we avoid saying more (and less) than what is needed. But it could be argued that it is at least sometimes possible to state a sentence that is fully compositional. It seems then that Fodor should accept that there are at least some sentences whose meanings can be obtained in a compositional way, and that language is potentially compositional, in the sense that sentences that are fully compositional could be stated, even if they generally aren't because they are not required for making ourselves understood.

Moreover, even in the case of sentences such as "it's three o'clock," it seems reasonable to

---

[114] Another way of trying to preserve the compositionality of language, also adopted by Elugardo (2005), is to say that Fodor's notion of compositionality is too strong. He denies that compositionality requires that the syntactic structure of sentences should be explicit about the structure of their contents. That is, he denies that there must be a one-to-one correspondence between the syntactic constituents of the sentence and the constituents of its semantic content, for the sentence to be compositional. But if one accepts that the content of the sentence "it's three o'clock" is that *it is three o'clock in the afternoon here and now*, and that the sentence doesn't have "here", "now" and "afternoon" as constituents, then the sentence's content is simply not being determined compositionally. Another way out is to say that the meaning of a sentence is the proposition, not the thought, that it expresses. In this case, however, one might also ask whether the proposition that a sentence expresses can always be derived compositionally from the parts of the sentence and their mode of combination. As Clapp notes, that would also be problematic. But I do not intend to explore these alternatives here.

[115] Recanati (2010), a contextualist, argues that pragmatic processes, which are not linguistically controlled, can enrich what a sentence says. For instance, the operation of modulation can alter the senses of the constituents of a sentence, depending on the context. But he then argues that compositionality can still be preserved, even if not all the constituents of the interpretation are made explicit in a sentence, provided that what determines the meaning of a sentence are the *modulated* senses of its constituents, together with the way they are combined.

suppose that their constituents and the way they are combined are not entirely irrelevant, but that they place some constraints on the possible meanings of the sentence. Fodor's argument does not seem to be enough to discard the idea that non-explicit sentences have their meanings, at least in part, determined by their constituents and the way they are combined. So if we are committed to Fodor's ideas that the meaning of a sentence is the thought it expresses, that compositionality requires explicitness of constituents, and that hidden constituents should not be multiplied with the only motivation being to preserve the compositionality of a sentence, then we might have to conclude that there is occasionally more in a thought than there is in a sentence that expresses it, and that not all sentences are strictly compositional. But we can still accept that some sentences are explicit and entirely compositional, that some sentences could be made explicit, and that even non-explicit, not fully compositional sentences, can have constituents that play some role in determining their meanings. That is, if we accept that there are degrees of compositionality, we can say that the meaning of a sentence is *at least in part* determined by the meanings of its constituents and the way they are combined, even in cases where the content of a sentence does not contain *all and only* what can be extracted from its constituents. A partial compositionality could explain the productivity and systematicity of language,[116] and allow us to preserve the productivity and systematicity arguments for the compositionality of thought (even if thought's compositionality, unlike the compositionality of language, is assumed to be complete because of thought's explicit character).

So far I have been dealing with Fodor's objection to the compositionality of language by granting him that the content of a sentence is the thought it expresses. I tried to show that, even if we accept this, one can still attribute at least some degree of compositionality to language. But another reasonable way of blocking Fodor's argument against the compositionality of language, is simply to deny that the content of a sentence is solely the thought that it expresses.[117] So one standard view assumes that there can be a separation of the *semantic content* of a sentence from its *assertion content*, which is the thought that a speaker intends to express in uttering a sentence on a

[116] Prinz (2012) uses a somewhat similar strategy, but with the purpose of supporting the view that some concepts are prototypes against the charge that prototypes don't compose. He argues that we are capable of combining concepts (and prototypes) compositionally, even if we don't always do so. According to him, potential compositionality is all we need to account for the productivity and systematicity of thought.

[117] This strategy is also adopted by Elugardo (2005) when challenging Fodor's argument against the compositionality of language.

particular occasion. The semantic content of a sentence could be compositional, determined by the conventional linguistic meanings of its constituents, even in a case where the assertion content of an utterance of the same sentence cannot be obtained in a strictly compositional way from the syntactic constituents of the sentence, perhaps because the sentence lacks constituents that the assertion content has. It could then be said that there is something any competent speaker of English understands from the sentence "it's three o'clock" in the absence of any particular context, which is determined compositionality from the sentence's structure and constituents. But the same sentence can have different assertion contents, depending on whether it is used to express the thought that it is three o'clock in the afternoon in Paris, or three o'clock at night in Chicago. Assuming the distinction between semantic content and assertion content is correct, Fodor's inexplicitness objection could at most be directed against the compositionality of assertion content.[118] For he could insist that the assertion content has constituents that are not present in the sentence. But nothing he has said speaks against the compositionality of semantic content. And perhaps all we need in order to explain the productivity and systematicity of language is the compositionality of semantic content.

We could question, though, the real need to assume that there is a kind of linguistic content, namely semantic content, that is purely compositional, in order to account for productivity. As we have seen, one of the main motivations for accepting the compositionality of language was to explain the productivity of language, our capacity to understand and produce sentences we never heard before. The idea was, for instance, that we, as hearers, derive the meaning of a new sentence solely from its parts and the way they are combined. I then claimed that if we were to accept Fodor's argument that language is not compositional, that would compromise the explanation we gave for the productivity of language. But, in fact, I believe that there was no need in the first place to assume that, whenever we hear a new sentence, its meaning is *solely* derived from the meanings of its parts and the way they are combined. There is no need to assume that sentences encode everything that we understand when we hear a sentence. We can assume that, usually, most, if not all, of the information is linguistically encoded. But there is no need to assume that what I understand from a sentence is *always* completely encoded in a linguistic manner. So if what we want to explain is our capacity to understand new sentences, there may be no need to assume that language is strictly compositional (despite the fact that this

[118] For a defense of the compositionality of assertion content against underdetermination arguments such as the one Fodor formulates, see Szabó (2010).

was the main motivation for saying that language is compositional). For we can say that what we understand, at least sometimes, is a combination both of the information that is encoded in the sentence, and of whatever other information is pragmatically available, in the extra linguistic context. As I suggested above, it may be enough to say, then, that language is partially compositional, in the sense that it encodes at least part of the meaning that is extracted in communication, but that it doesn't *have* to encode everything that is being communicated. If that is all we need to account for productivity, then we might not even have to assume that there is a form of content (namely semantic content) that is strictly compositional – at least not for the purpose of explaining productivity.

Even if we don't really need to assume that language is strictly compositional in order to explain its productivity, we still might want to say that the meanings of the constituents of a sentence and the way they are combined impose some constraints on the possible interpretations of the sentence. And compositionality may still be what explains its systematicity. Also, as I tried to show, in denying that language is compositional, Fodor compromises the arguments for the language of thought presented in the first two sections of this chapter. These consequences could perhaps be tolerated if the non-compositionality of language could be used to solve the thought-or-language problem, that is, if it could be used to determine that it is thought, not language, that has meaning in the first place. But, as we have seen, Fodor does not reach this conclusion in a neutral way, for his argument is circular. Again, Fodor cannot claim that the non-compositionality of language is what is deciding that language does not have content in the first place, because in order to argue against the compositionality of language, he is already assuming that natural language sentences derive their meanings from the thoughts they express.

Since compositionality cannot really decide the thought-or-language problem, it is better to hold either to a mitigated compositionality of language or to a compositionality of the semantic content of sentences. That way, we can at least partly explain the productivity and the systematicity of language, which seems quite difficult to deny, and we can preserve the arguments for the language of thought which take as their basis the analogy with language. The admission of some degree of compositionality in language would possibly prevent Fodor from following his plan to use compositionality as a criterion for deciding whether it is thought or language that

comes first in the order of explanation.[119] But, as we have seen, this is a plan that cannot really be followed in a non question-begging way.

This is not to say that there is no reason to assume that the content of language is, at least sometimes, derived from the content of thought. On the contrary, that is an intuitive assumption, and it is behind the very idea suggested above, that the same sentence can, when used in different situations, have different assertion contents. Also, non-literal meanings of words and sentences are often explained in terms of the intentions of the speakers in uttering those words and sentences. More generally, as we are going to see in the next chapter, unlike language, thought is often assumed both to be unambiguous and also what disambiguates sentences and words in a natural language (Pinker 1994, Pylyshyn 2003, Fodor 1998). So some phenomena can be explained by the assumption that sentences and words mean what they do because they express the thoughts that they express. But the compositionality of thought and the alleged non-compositionality of language, as I've tried to show, are consequences of the assumption that thought, unlike language, has underived content. They do not, then, give us any extra reasons to assume the priority of thought over language. This is why compositionality doesn't really solve the thought-or-language problem. It is possible that whatever is prior is thereby compositional. Once we assume that thought has underived content, it may be natural to assume that it is compositional. But this may be an argument for the compositionality of thought, and not for its priority. In the next chapter, we will consider some reasons for assuming the priority of thought over language.[120]

---

[119] In this new scenario, in which we accept that there may be degrees of compositionality in language, Fodor could still say that only what is *strictly* compositional has content in the first place. But he would then have to argue against the strict compositionality of language, and in favor of the strict compositionality of thought, without presupposing the priority of one with respect to the other, that is, without presupposing that the content of a sentence is the thought it expresses, and that a thought just is its content.

[120] A shorter version of this chapter will be published in *Philosophical Papers*. I thank two anonymous reviewers for helpful comments.

CHAPTER 4

LANGUAGE OF THOUGHT AND NATURAL LANGUAGES

We have seen in the previous chapter the productivity and systematicity arguments for the idea that the system of mental representations is a language. In this chapter, the goal is to explore how the language of thought is supposed to be related to the natural languages. In the first 4 sections, I will briefly present some of the possible views on the relation between thought and language, including the main arguments that some of the proponents of the language of thought have formulated for the idea that it must be different from any natural language. My general goal is to investigate, within the language of thought paradigm, whether the natural language we speak can still have some influence on the concepts we have. In order to do that, I will first explore some of Fodor's views about concepts – such as their innateness and structure (section 5) –, as well as his views about the relationship between the semantics of thought and the semantics of language (section 6). In section 8, I will discuss the issue of the variation of color terminology across languages, as well as some empirical findings about the influence of color words on non-linguistic cognitive domains. Color words and concepts offer a good case study because their mutual influence has been relatively well studied by linguists and psychologists, though rarely with any attempt to clarify more general philosophical questions. Finally, in sections 9 to 14, I will criticize some of Fodor's views, for they don't seem compatible with the empirical evidence that at least some of our concepts vary according to the language we speak. I think Fodor may be right that most of our lexical concepts are primitive, that is, not internally structured, but not that being primitive implies being innate. I also think he is too quick to reject the idea that the semantics of natural language can influence the semantics of thought. The view I will sketch is one according to which the language we speak is not the vehicle of thought, but can influence the semantics of thought, in that it causes some of our concepts to have the extensions that they have.

1. Thought in nonlinguistic creatures

> "In an earlier age, the absence of language was used as an argument against the existence of thought in other species. Today I find myself upholding the position that the manifest reality of thinking by nonlinguistic creatures argues against the importance of language."
>
> (Frans de Waal)

There are several different questions that we can ask regarding the relation between thought and language. One that has long interested philosophers is whether there can be thought without language or whether language is necessary for thought. A natural way to try to approach this issue is to consider whether non-linguistic creatures are capable of thinking. Descartes, for instance, held that lacking language is a sign that animals are incapable of thought.[121] Descartes probably didn't think that language was necessary for thought, in the sense of language being what makes us capable of thinking, but rather he thought that language was an expression of thought, and our main evidence for it. As he puts it, "speech is the only certain sign of thought". Since animals lack language, Descartes took this to be an indication that they lack thought as well. According to him, the behavior of animals, unlike that of humans, could be explained in a purely mechanistic way. Language is a sign that humans have reason, which, according to him, could not be mimicked by machines (see DESCARTES, 1637, p. 46-7).

Of course, much depends on what we mean by "thought" and "thinking",[122] and it is not my intention to explore Descartes' views. But contra Descartes, there is now a growing consensus that (at least some) animals and pre-linguistic infants exhibit sophisticated behavior, even if non-linguistic, which is explainable only if we attribute complex cognitive capacities to them. As

[121] As Descartes says in a letter, "in my opinion the main reason for holding that animals lack thought is the following. Within a single species some of them are more perfect than others, as humans are too. This can be seen in horses and dogs, some of which learn what they are taught much better than others; and all animals easily communicate to us, by voice or bodily movement, their natural impulses of anger, fear, hunger, and so on. Yet in spite of all these facts, it has never been observed that any brute animal has attained the perfection of using real speech, that is to say, of indicating by word or sign something relating to thought alone and not to natural impulse. Such speech is the only certain sign of thought hidden in a body. All human beings use it, however stupid and insane they may be, even though they may have no tongue and organs of voice; but no animals do. Consequently this can be taken as a real specific difference between humans and animals." (1649, p. 366). He says also that "there are no men so dull-witted and stupid, not even madmen, that they are incapable of stringing together different words, and composing them into utterances, through which they let their thoughts be known; and, conversely, there is no other animal, no matter how perfect and well endowed by birth it may be, that can do anything similar." (1637, p. 47).

[122] And also on what we mean by "language". But let's assume for now that animals lack language, whatever exactly language is.

Carruthers notes, animals "can decide whom to form an alliance with, or can calculate rates of return from different sources of food, or can notice and exploit the ignorance of another" (2002, p. 661). To give just a few examples, corvids, like apes, act in anticipation of future events (KABADAYI; OSVATH, 2017), and evidence suggests that monkeys, as well as apes, understand complex social relations among the individuals of their groups (CHENEY; SEYFARTH, 1990).[123] Summarizing the findings of several studies, Tomasello and Herrmann say that

> apes not only perceive and understand things in the immediate here and now but they also recall things they have perceived in the past and anticipate or imagine things that might happen in the future. (…) Great apes also can make inferences about what one perceived state or event implies about another. (…) Apes also can reason about the decision making of other individuals. (TOMASELLO; HERRMANN, 2010, p. 04).

There is now even evidence that apes can attribute false beliefs to others, which was thought to be something uniquely human (KRUPENYE et al., 2016). Attribution of cognitive states to non-linguistic creatures is now the paradigm in ethology, as well as in the study of pre-linguistic human infants, without which it is hard to explain some behaviors.

It seems, then, that language is not our only indication of thought, or of representational states and cognitive processes, though it is certainly one of them. And if animals think, or have cognitive capacities, then language is not necessary for all thought – though it may be sufficient for it. At the heart of psychological or cognitive explanations of behavior is the attribution of representational states to individuals. Theorists may disagree about the nature of the representational vehicle employed by non-linguistic creatures. Some will say that animals deploy mainly a perception-based type of representation. Others, like Carruthers, embrace the language of thought hypothesis – the idea that complex thoughts are composed out of concepts, which are symbols that don't need to be couched in perceptual or natural language representations. According to him, "it is well-nigh impossible to see how apes can be capable of representing multiple, complex, and constantly changing social relationships (who is friends with whom, who

[123] "Monkeys in many different species appear to observe interactions in which they are not involved and recognize the relationships that exist among others. In this respect, monkeys make good primatologists. A male considers how strongly a female prefers her partner before he attempts to take her away; juveniles and adult females take note of their opponents' kin as they plot retaliation or reconciliation; and adult females, upon hearing a juvenile's cry for help, learn to expect a response from the mother." (CHENEY; SEYFARTH, 1990, p. 175).

has recently groomed whom, who has recently fallen out with whom, and so on) unless they are capable of structured propositional thought." (CARRUTHERS, 2002, p. 662).[124]

Carruthers has gone as far even as to argue that bees have conceptual thought. A lot has been written about bees' navigation systems, and scientists commonly explain their capacity to navigate in the environment by saying that bees represent their location and the location of the food source by means of mental maps. But even if we accept this explanation, can we infer from this that bees have concepts? What are the requirements that need to be fulfilled in order for a creature to have conceptual thought? As Carruthers notes, one constraint on concept possession, proposed by Gareth Evans (1982), is the *generality constraint*. As Carruthers interprets Evans, the constraint is that

> genuine thinkers must be capable of entertaining all syntactically permissible combinations of any concepts that they possess (…). So if thinkers possess the concepts *F*, *G*, *a*, and *b*, then they must be capable of thinking each of the thoughts *Fa*, *Fb*, *Ga*, and *Gb* (but not the "thoughts" *FG* or *ab*, which are ill-formed and uninterpretable). (CARRUTHERS, 2009, p. 94).

According to this view, having thoughts with conceptual content requires being capable of combining the more basic representational units that one has in all permissible ways.[125]

Carruthers notes that it is too strong a requirement that, in order for a creature to count as a concept user, it should be capable of combining *all* of its concepts in *all* permissible combinations. In fact, it is unclear what motivates such a strong constraint, since it is questionable that even humans satisfy it. For instance, we have the concepts SQUARE, NUMBER, GREEN, ROUND, LOUD.[126] But one could conceivably argue that we can't really combine them in all syntactically permissible ways. We can't really think ROUND SQUARES, or LOUD GREEN

---

[124] Evidence also suggests that earlier hominids were also capable of sophisticated thinking. As Carruthers points out, "*Homo erectus* and archaic forms of *Homo sapiens*, for example, were able to survive in extremely harsh tundra environments, presumably without language (…). It is hard to see how this could have been possible without a capacity for quite sophisticated planning and a good deal of complex social interaction." (CARRUTHERS, 2002, p. 662).

[125] It is unclear whether, for Evans, this was supposed to be a constraint that applies only to human thinking, or to thought in general. As he puts it, "if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception. This is the condition that I call 'The Generality Constraint'." (EVANS, 1982, p. 104) For a detailed discussion of Evans' use of the generality constraint, and of how it contrasts with the use that other philosophers, such as Carruthers, make of it, see Clapp and Duhau (2011).

[126] Here I follow the convention of using uppercase for names of concepts.

NUMBER (though we can certainly *say* round square). Carruthers proposes a weaker formulation of the constraint, according to which "if a creature possesses the concepts *F* and *a* (and is capable of thinking *Fa*), then for *some* other concepts *G* and *b* that the creature could possess, it is metaphysically possible for the creature to think *Ga*, and in the same sense possible for it to think *Fb*." (2009, p. 97). Much more reasonable, then, would be to adopt this weaker requirement, that if a creature possesses concepts, it is in theory capable of combining them with at least some of the other concepts that it has.[127]

Back to bees. Carruthers thinks they have a representational system that satisfies the weaker generality constraint, even if it fails to satisfy the stronger constraint. He argues that the map-like representations that are often attributed to bees – in order to explain their capacity to navigate to and from a food source – have simpler elements that can be recombined. They use *symbolic* maps, whose symbols can be combined in different ways. According to him, a bee that can think a

> thought with the content [nectar is 200 meters north of the hive] (or some near equivalent) (…) can also think thoughts with the contents [the hive is 200 meters north of the nectar], [nectar is 200 meters west of the hive], [pollen is 400 meters north of the hive], and so on for all interpretable combinations of the four candidate concepts, both with each other and with other similar concepts. And we know that the inferences in which bees engage are sensitive to such component structures. (CARRUTHERS, 2009, p. 98).

Carruthers note, though, that bees might be incapable of thinking that there is nectar 200 meters north of the brood chamber, even if they are capable of employing each of these concepts in other thoughts. According to him, that's because bees might use inside the hive a different representational system than the one they use outside the hive, to locate food. But even if that's the case – even if some combinations of concepts cannot occur for bees –, it is unclear why we should deny that they have concepts on the basis of the stronger generality constraint. Assuming that they represent the world by means of representations that can occur in different

[127] Perhaps humans come closer to satisfying the stronger constraint, but if that's the case, it is presumably not because we possess concepts and animals don't. It is not possessing concepts that in itself gives us the capacity to combine them in all kinds of ways. Carruthers hypothesizes that humans have a capacity for supposition and creative thinking that is not shared by other animals, which is responsible for us coming closer to satisfying the stronger formulation of the generality constraint. I can't say whether or not Carruthers is right about that supposition, but I agree with him that the requirement that a creature be able to combine concepts in all possible ways is too strong a restriction on concept possession, and it seems unmotivated.

combinations, albeit not all possible ones, why should we take these representations not to be really conceptual simply because they don't satisfy the stronger constraint (that is, because their combinatorial possibilities are restricted)?

It is worth making a quick digression to note that the generality constraint proposed by Evans is similar to what Fodor calls systematicity (a notion that we've seen in chapter 3), but systematicity plays a different role in Fodor's views. Fodor doesn't use it as a constraint on concept possession. According to Fodor, thought seems to be systematic, in the sense that the capacity to think certain thoughts is associated with the capacity to think other, related thoughts. It is a fact that anyone who can think, say, that John loves Mary, can also think that Mary loves John. Fodor thinks this fact can be explained if we assume that thoughts are structured entities, composed out of simple elements (concepts) that can be combined in different ways. But Fodor doesn't think systematicity is a constraint on conceptual thought. He is not saying that only creatures that exhibit systematicity can count as creatures that are capable of thinking. He is simply saying that human thought (and presumably the thought of some animals) is frequently systematic, and that is explainable by the assumption that thought has a constituent structure. Moreover, it is compatible with the observation that thought is systematic that not all concepts can be combined with all others. If we found that we can combine the simple concepts *a*, *b*, *F*, *G*, and *H* into the thoughts *Fa*, *Gb*, *Ha* and *Hb*, while being incapable of thinking the thoughts that *Fb* and *Ga*, that wouldn't show that thought is not systematic. The point of the systematicity argument is simply that it is reasonable to infer, from a creature that exhibits systematicity to some degree, that that creature has thought with a constituent structure. But it's a different thing to say, as the generality constraint does, that *only* creatures that have a completely systematic thought can count as creatures that have concepts.

In fact, Fodor's requirements for concept possession are much weaker. For him, one possesses a concept if one is capable of thinking about whatever property or individual that concept refers to. I posses the concept RED if I'm capable of thinking about redness (cf. 2008, p. 48). "Having a concept is being able to bring to mind whatever it is the concept of" (FODOR, 2003b). So I don't necessarily need to be capable of thinking about red sneezes, or red numbers, in order for me to count as a creature that possesses the concept RED. Generality constraint and systematicity are related notions, but they are not identical, and they have different functions in philosophical discussions. The generality constraint is a constraint on concept possession, that is,

on what is required for a creature to have concepts. It states that a strong form of systematicity is a requirement for concept possession. Someone like Fodor, though, could say that a creature's thought is systematic, and infer from this that that creature has conceptual representations, without being committed to the idea that *only* creatures with strongly systematic thoughts have concepts. If Fodor were convinced that bees' thoughts are systematic, he would probably infer that they have conceptual thought. But he would not, I think, infer that bees don't have conceptual thought in case they don't have thoughts that are strongly systematic (allowing all possible combinations of constituents).

Whether or not Carruthers is right in thinking that bees have conceptual thought,[128] the fact is that attributing representations to animals, especially conceptual representations that combine to form structured thoughts, helps us explain some of their behavior in ways that purely mechanistic or biological explanations don't. It is because bees represent spatial relations among objects of interest that they can navigate their environment – and even communicate to others the location of sources of food. It is because apes represent certain individuals of their group in a way that differs from the way they represent others, and because they represent the relations among individuals in their group, that they behave the way they do when engaging in social interactions. It is because apes expect that a person has a false belief about the location of an object that they look at the place where that person last saw an object when that person reenters a scene, even though the ape knows that the object is no longer there. Animals use these representations in interacting with others, in inferences while solving problems, in planning their behavior, etc. It is doubtful that we could explain all there is to explain about animal behavior without appeal to representations.[129]

If the appeal to representations proves fruitful in psychological explanations, as it has so far proven itself, then that suggests we are at least provisionally committed to the assumption that there are mental representations, and that animals, including humans, employ them in thinking. I

---

[128] For a criticism of the idea that fulfilling the weaker generality constraint is sufficient for concept possession, see Duhau (2010).

[129] In an interesting study by Lin et al. (2007), the researchers found that three types of individual neurons in the hippocampus of mice respond specifically when they encounter nests or beds. The response is invariant regardless of the nest's location, shape, color, odor and material. According to them, the study reveals the neural mechanisms that encode the abstract concept NEST. They are certainly following the reasoning I'm calling attention to here. They assume that mice's ability to recognize nests is a sign that they have an abstract concept of nests, which makes that recognition possible and explains the mice's behavior. If that's the case, we should be able to find neurons that respond specifically to nests – which is what they found.

leave open the possibility that purely biological explanations might take over and render some psychological explanations of behavior unnecessary. But, as Bermúdez notes,

> seems highly unlikely (…) that satisfactory explanations in wholly nonpsychological terms will be found for all the behaviors for which psychological explanations are canvassed. It is difficult to imagine that researchers into animal behavior, infant development, and human evolution will return to the methodological precepts of behaviorism. (BERMUDÉZ, 2003, p. 07).

We can ask whether we need to appeal to representations in order to explain *all* kinds of behavior, or whether sometimes that appeal is unnecessary. We can ask whether those representations are abstract (amodal) symbols, or whether they are imagistic (perceptual), or whether animals and humans use perhaps both kinds of representations. We can also ask what their content is, and what makes a representation have the content that it has. We can ask whether we can know precisely the vehicle and content of the representations employed by animals and prelinguistic infants. But from now on I'm going to take it for granted that we need representations to explain at least some kinds of behavior.[130]

I don't expect these considerations to convince the most skeptical of philosophers, but I propose to take seriously the growing consensus in the scientific community, and to assume that at least some cognitive processes occur without natural language. Language is not necessary for all thought. In many circumstances, animals behave in ways similar to our own. Assuming that psychological states and processes are part of what explains human behavior, the simplest explanation here is that animals behave the way they do because they also have a rich mental life – and not because they are machines deprived of thoughts and feelings. It is simpler, then, to assume that the different kinds of behavior that animals exhibit are explained under the assumption that they have beliefs and desires (which are intentional, or representational states), as well as cognitive processes like memory, perception, categorization, reasoning, planning, etc., which are processes that involve the transformation of representations. Human behavior is no doubt a lot more sophisticated than the behavior of any other animal. But the differences between

[130] I take that the systematicity and productivity arguments presented in chapter 3 are good arguments for the existence of representations with symbolic structure. To the extent that animals' thoughts are productive and systematic, we would have good reasons to assume that they are, like ours, symbolic.

humans and other animals are unlikely to be in the capacity to represent, or to think about the world. Animal behavior suggests that having concepts, and deploying those concepts in inferences, is not what distinguishes humans from other animals. If that's the case, then language is not necessary for the existence of conceptual thought.

2. Communicative vs. cognitive conceptions of language

It will be useful to introduce here a distinction made by Carruthers (1996, 2002) between natural language being *required* (or being a *prerequisite*) for certain types of thought, and language being *involved* in certain types of thought as its *medium*, or *vehicle*. So one way language can affect thought is by being required for us to be able to think certain thoughts. Another way language can affect thought is by being the very vehicle of some thoughts. I indicated in the previous section that there is evidence that animals have cognitive processes, or thoughts. If that is correct, and assuming that animals lack language, then the view that language is necessary for all thought is false. It is, in fact, false both that all thought requires language and that all thought involves language as its medium.

To emphasize this point, Pinker (1994) brings up other examples of organisms able to think without language: deaf people who live to adulthood without learning any sign language, and aphasics, individuals who have the abilities to speak and/or understand language compromised due to brain damage. According to him, all these individuals prove to be able to have different types of non-verbal thoughts, which again suggests that language is not necessary for all thought.[131] Fodor similarly uses the evidence for thought in animals and pre-linguistic infants to argue against the idea that natural language is the medium of thought. According to him, "the obvious (and, I should have thought, sufficient) refutation of the claim that natural languages are the medium of thought is that there are nonverbal organisms that think" (FODOR, 1975, p. 56).

But even if language is not necessary for some kinds of thought to occur, it doesn't follow that no thought requires or involves language. For it could still be the case that language is the

[131] It should be noted, however, that we cannot infer, from the case of aphasia alone, that language was not required for the cognitive processes that are found to be intact in those individuals. It could be that individuals with aphasia can still perform well on some cognitive tests and behave normally in some respects because they learned a language at some point – language being a prerequisite for the types of thought that they exhibit.

medium of, or at least a prerequisite for, some types of thought, which don't occur in animals, infants and linguistically impaired adults. In addition, from the fact that natural language is not required for animals or pre-linguistic infants to perceive, remember, categorize, learn and plan for the future, and therefore not involved in these cognitive processes, it also doesn't follow that in our own case, as adult linguistic beings, natural language is never the medium of any of these or other cognitive processes. So from the fact that there can be thought without language, it doesn't follow that natural language is dispensable for all kinds of thought, nor that it is not involved in any thoughts.

As Carruthers sees the different positions on the relationship between thought and language, there are those who adopt a *communicative* conception of language, and those who adopt a *cognitive* conception of language.[132] What really distinguishes these positions, according to him, is that those who think that language serves merely communicative purposes would deny that language is ever the medium of thought. Proponents of the cognitive conception, on the other hand, would argue that natural language is the medium of some types of thought, and therefore that it has, besides a communicative function, also a cognitive function. A proponent of the communicative conception would say that we use language mainly to express thoughts, whereas a proponent of the cognitive conception would say that we also use language to think.

Both proponents of the communicative conception of language and of the cognitive conception would agree, says Carruthers, that it is only because we learned a language that we can come to think about, say, voting, atoms, the planet Saturn or the number 479,567,904. So no one denies that language is an extremely important feature of our lives. There are not only things about which we think that we wouldn't think about had we not learned a language, but there are also all sorts of things that we create that could hardly have been created had we no means to communicate with others. Without language, we most likely wouldn't have created agriculture, sophisticated societies, airplanes, telescopes and vaccines; we wouldn't have deciphered the structure of DNA or landed men on the moon. For better or worse (mostly better), there would be no PhD dissertations in philosophy. Language seems to be de facto necessary for the acquisition

[132] This is not to deny that language can also be used to perform actions, as Austin noted. A priest marries a couple by saying certain words and not others. I make a promise to someone by saying "I promise to do X". It could perhaps be said that in performative speech acts we communicate certain intentions – in which case, language's primary function continues to be the communication of thoughts, with the additional feature of constituting certain actions. But I don't intend to deal with this discussion here.

of many concepts we regularly employ.[133] As Carruthers (2002) sees the issue, no one denies the *diachronic* importance that language has for us being capable of thinking certain thoughts. What he thinks is interesting to investigate is whether language has also a *synchronic* influence on thought. So it is one thing to say that language is an important vehicle for *communicating* ideas and building knowledge, and another that it is *involved* in certain or all of our thoughts. A proponent of the cognitive function of language would say that natural language is constitutively involved in at least some of our thoughts. A proponent of the communicative function of language, on the other hand, could agree that we come to think about many things that we wouldn't otherwise have thought about because we learned them through language. So he would agree that language is necessary, or is a prerequisite, for the acquisition of some concepts. But would say that that doesn't necessarily mean that language is actively involved in any thoughts as their vehicle.[134]

On the extreme end of the cognitive view, we can place the proponents of what is now known as linguistic determinism. According to this view, the particular natural language we speak determines the way we think. In two often-quoted passages, Whorf says that

> the background linguistic system (in other words, the grammar) of each language is not merely a reproducing instrument for voicing ideas but rather is itself the shaper of ideas. (…) We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds — and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way — an agreement that holds throughout our speech community and is codified in the patterns of our language. The agreement is, of course, an implicit and unstated one, BUT ITS TERMS ARE ABSOLUTELY OBLIGATORY; we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees. (WHORF, 1940, p. 213-14).
>
> We are thus introduced to a new principle of relativity, which holds that all observers are not led by the same physical evidence to the same picture of the

[133] I think it is conceivable that one could have thoughts about, say, DNA and atoms without ever having learned a language (Carruthers envisions a thought experiment to show this in 1996, p. 21). Assuming conceivability entails possibility, it is possible to have these thoughts without ever having learned a language. But this sense of possibility is that of metaphysical possibility, and not of nomological possibility. In the world as we know it, with the psychology that we have, that doesn't seem to be a real or nomological possibility.

[134] Although Carruthers says that proponents of the communicative conception of language accept that language is a prerequisite for our acquisition of certain concepts, in section 7,I will argue that it isn't clear that Fodor accepts that.

> universe, unless their linguistic backgrounds are similar, or can in some way be calibrated. (WHORF, 1940, p. 214).

This suggests that, prior to the acquisition of a natural language, we don't really have any concepts, we don't really organize the world in any particular way. Language has the cognitive function of basically dividing the world into categories. And thought will differ among individuals as much as the languages they speak differ. That surely is an extreme view. For one thing, it leaves unexplained the cognitive capacities that are exhibited by animals and pre-linguistic children. If we used language to think, in a strong sense, if thought was just an amorphous mass, as Saussure puts it, before language acquisition, it would be difficult to explain how animals and infants manage to survive in a world where food is no different from stones.

So for the linguistic determinist we not only use language to think, language being the vehicle of our thoughts, but we also would not be able to think at all if we hadn't learned a first language. Language is necessary for thought in a strong sense.[135] Descartes, as we've seen, also held an extreme position, but as I read him, he could still be considered a proponent of the communicative conception of language. He held that language is a sign or *expression* of thought, which, as he claims, occurs even in deaf people who, though lacking the phonological input, still have the impulse to communicate. Descartes says nothing about natural language being constitutively involved in thinking as its vehicle. Fodor, Pinker and Pylyshyn are also proponents of the communicative view of language, for they hold that language serves only the purpose of communicating our thoughts, and they think that it is not constitutively involved in our thoughts. Unlike Descartes, though, they don't infer the lack of thought in animals from the lack of a language to express it. Carruthers, on the other hand, is a proponent of the cognitive conception of language, and he holds that natural language is involved in conscious propositional thoughts. In the next section I present some of the main arguments from proponents of LOT against the idea that natural language is the vehicle of thought.

---

[135] The issue is in fact a little more complicated than that. It is not completely clear that Whorf would say that we use language as a medium for thought, and not simply that language initially molds our concepts or ideas, being later merely used to express them. I in fact think that Carruthers' characterization of the opposition between communicative and cognitive conception of language, though useful, is too restrictive. For one thing, there are positions in the logical space that would seem to deserve the label "cognitive", while not being clearly committed to the idea that natural language is the vehicle of thought (for instance Whorfian forms of linguistic relativism). I will explore this view in the final sections of this chapter.

3. Against introspection

> "I am not sure that I think in words, and I never seem to hear any inner voices." (Frans de Waal)

A natural reaction against the idea that natural language serves primarily a communicative purpose, as opposed to having any serious cognitive function, is that it conflicts with what introspection tells us about thinking (see Carruthers, 1996). It seems undeniable that we often experience our thinking as something like an internal monologue or inner speech – which occurs in a particular natural language, typically our native language. Carruthers has proposed that we should take the experience of inner speech seriously. For him, natural language is the medium of conscious propositional thinking. It is true that some people, notably Temple Grandin (1995), claim that their conscious thoughts occur predominantly in visual images. Pinker (1994) also notes that some physicists and mathematicians claim to think geometrically, not verbally, while doing calculations and solving problems. But Carruthers, like Fodor, thinks that visual images are not good candidates for being the vehicle of propositional thought, for there is often a one-to-many relation between an image and its possible contents. The same image can be used to stand for several different propositions. But whether we primarily introspect our own thoughts as occurring in sentences in a natural language or in visual images, we don't seem to introspect our own thoughts as occurring in a non-natural language. Why then assume the existence of a language of thought which is different from anything that is given to us through introspection?

As I indicated a few paragraphs back, Fodor and Pinker argue against the idea that natural language is necessary for thought by saying that animals and pre-linguistic infants think. But even if we accept that they are right about this, as I think they are, that in itself doesn't show that natural language is never the medium of at least some thoughts of competent language users. And what Carruthers is claiming is that introspection shows that natural language is the medium of at least our conscious, propositional thoughts. But proponents of LOT (acronym for *language of thought*) have formulated several other arguments against the idea that natural language or mental images are the medium of thought, and consequently against the idea that introspection should

be taken as a reliable indicator of what the vehicle of thought is. I will briefly mention here some of the arguments formulated by Fodor, Pylyshyn and Pinker.

Perhaps the main problem that has been raised against the idea that we think in sentences in a natural language is that natural language sentences are quite often ambiguous. Thoughts, on the contrary, seem to have a determinate meaning – in fact, thought seems to be what disambiguates sentences in a natural language. So thoughts cannot be identified with the sentences that are used to express them. As Fodor says in *LOT 2*, "that there are ambiguities in English is, indeed, the classic reason for claiming that we don't think in English" (FODOR, 2008, p. 73). Consider the English sentence "I'll meet Sarah at the bank". It seems that it can be used to express several different thoughts – "Sarah" can be used to refer to several different people named Sarah, and "bank" can be used to refer to the financial institution, or the river area. But if this sentence occurs to me in inner speech, and it plays some role in my thinking and acting, it is reasonable to assume that it is given a definite meaning, with the name "Sarah" expressing the concept I have of a particular person I know named Sarah, and "bank" standing for a concept that refers either to the financial institution, or to the river bank. I don't have the problem of not knowing who I mean by "Sarah", or what I mean by "bank", which an interlocutor might have if I were to say that sentence out loud. So thought seems to be free from the lexical ambiguities that permeate natural languages. Whereas the same sequence of sounds can be used to refer to completely different things, one concept will never refer to both financial institutions and the river areas, for concepts are partly individuated by their referents, and these are clearly two distinct entities in the world.

Thought is also free from syntactic and scope ambiguities. For instance, the English sentence "Every man loves a woman" can express the thought that for each man there is some woman or other that he loves, but can also express the thought that there is a particular woman such that every man loves her. So, if the sentence "every man loves a woman" occurs to me in inner speech while I'm thinking, and it is reasonable to assume that I'm not merely saying things in my head, then it occurs to me as an expression of a determinate thought. Given that the mere sequence of words in English could be the same for either of those thoughts, it's reasonable to assume that there must be more to thought than inner speech in a natural language. As Fodor puts it,

> 'thinking in English' can't just be thinking in (sequences of) English words since, notoriously, thought needs to be ambiguity-free in ways that mere word sequences are not. There are, for example, two thoughts that the expression 'everybody loves somebody' could be used to think, and, so to speak, thought is required to choose between them; it's not allowed to be indifferent between the possible arrangements of the quantifier scopes. That's because, sans disambiguation, 'everybody loves somebody' doesn't succeed in specifying something that is susceptible to semantic evaluation; and susceptibility to semantic evaluation is a property that thought has essentially. You can, to be sure, say in your head 'everybody loves somebody' while remaining completely noncommittal as to which quantifier has which scope. That just shows that saying things in your head is one thing, and thinking things is quite another. (FODOR, 1998b, pp. 64-5).

So thoughts that are expressed by the same ambiguous sentence represent different things, and these differences must be explicit in their vehicles. Besides, they have very different causal roles. If we want to explain their different causal powers, it helps assuming that they are different thoughts, despite them being expressed in the same way.

Pylyshyn raises a similar problem against the idea that we think in a natural language. According to him, sentences very often do not express everything that we mean when having a given thought. Pylyshyn asks us to consider a banal example of a conscious thought of a sentence that occurs to him during the writing of a text: "I'd better hurry and finish this section or I will be late for my meeting." According to him, there are a number of things he knows, but that are not explicit in the sentence. For example, he knew which meeting he had in mind when he only thought "my meeting". He also knew how much time was meant by "late", what counted as "hurrying", and to whom the word "I" referred to. But if all that he knows relative to the meaning of the sentence is not expressed by the sentence, then all that knowledge must be encoded in thought in a different form, and not in the form of a sentence in a natural language (PYLYSHYN, 2003, p. 430). The sentence that is given through introspection is underdetermined. Much like mental images, it could be used to express several different thoughts, but the thought that Pylyshyn actually has is well determined, so the thought cannot be identified with the sentence given to introspection.[136] According to Pylyshyn,

[136] Both Fodor and Pylyshyn argue against the idea that mental images might be the vehicle of thought. According to them, images suffer from similar problems to those of sentence: they can be ambiguous like sentences, so that the same image can be used to represent different thoughts. Also, as we've seen in chapter 3, the productivity and the systematicity of thought are explained by the assumption that thought is like a language, in the sense of having syntactic structure and compositional semantics. Neither Pylyshyn nor Fodor deny that some people introspect

> What one is aware of – the form and content of one's conscious thoughts— cannot be what plays the causal role in reasoning. (…) The sorts of things of which one is aware, such as the words or sentences of an "inner dialogue" or the "mental pictures" one imagines, greatly underdetermine what one is thinking at the time one has those experiences. Consequently, something other than more words or more images must be going on. (PYLYSHYN, 2003, p. 430).

Fodor holds a similar view in the paper "Language, thought and compositionality." He thinks it is clear that, in general, a sentence does not have the same constituent structure as the thought that it is used to express. As we have seen in chapter 3, he gives as an example the sentence "it's three o'clock" when uttered as a response to the question of what time it is. According to him, what one intends to communicate, in the usual case, is that it is three o'clock in the afternoon here and now, but the sentence is inexplicit about the time of day and location. Since, according to him, thoughts are always explicit about their contents, but sentences are usually inexplicit about the thoughts they express, thoughts and natural language sentences cannot be identical.[137,138]

Another reason Pylyshyn gives against the idea that we think in a natural language is that there are concepts for which there are no corresponding words. But he also claims that we can have thoughts about perceptual experiences whose contents *cannot* be expressed in words. I can, for example, think that *this* is the same color as *that*, where the colors given to perception are the contents of my thoughts. Since the contents of some thoughts about perceptual experience needn't be categorized, Pylyshyn says that "some thoughts (…) can contain unconceptualized

---

mental images, but they deny that images have an important role in our reasoning. But I do not intend to enter the debate on mental imagery.

[137] In fact, what Fodor wants to show with this observation is ultimately that natural languages cannot be compositional, as we've seen in chapter 3. But it is implicit in his argument that, because thoughts are explicit and sentences are not, natural language sentences cannot be the vehicle of thought.

[138] Pinker makes basically the same point, when he says that "sentences in a spoken language like English or Japanese are designed for vocal communication between impatient, intelligent social beings. They achieve brevity by leaving out any information that the listener can mentally fill in from the context. In contrast, the 'language of thought' in which knowledge is couched can leave nothing to the imagination, because it *is* the imagination. Another problem with using English as the medium of knowledge is that English sentences can be ambiguous. When the serial killer Ted Bundy wins a stay of execution and the headline reads 'Bundy Beats Date with Chair,' we do a double-take because our mind assigns two meanings to the string of words. If one string of words in English can correspond to two meanings in the mind, meanings in the mind cannot be strings of words in English. Finally, sentences in a spoken language are cluttered with articles, prepositions, gender suffixes, and other grammatical boilerplate. They are needed to help get information from one head to another by way of the mouth and the ear, a slow channel, but they are not needed inside a single head where information can be transmitted directly by thick bundles of neurons. So the statements in a knowledge system are not sentences in English but rather inscriptions in a richer language of thought, 'mentalese.'" (PINKER, 1997, p. 70).

contents" (2003, p. 432) that are not expressible in language. One could argue against Pylyshyn that, in using demonstratives, that is precisely what we are doing: expressing perceptual contents through language. But I take his point to be that thoughts about perceptual experiences are much richer, more fine-grained, than words. I don't need to have a category for the particular shade of gray that my laptop has right now on its left side, but I'm capable of thinking about it, and of comparing it to the gray I see reflected on my mug. Perhaps because I have no such category, I have in my vocabulary no word that refers precisely to this shade of gray. But that doesn't prevent me from being able to think about this shade of gray. So I am likely thinking about this shade of gray by non-linguistic means. I can certainly refer to this color by the use of a demonstrative, whose content is determined by the context. But again, we seem to be facing a problem of the underdetermination of content by words: the same demonstrative, "that", can have countless different contents depending on its context of use; they are not specified in the word itself. And while demonstratives allow me to communicate my thoughts about perceptual experience, the demonstratives themselves say very little about the content of those experiences.

So although introspection may lead some of us to believe that sometimes thoughts occur in the form of an inner speech, for the reasons presented, Pylyshyn concludes that thoughts cannot be identified to sentences that are given to consciousness.[139] According to him, "the things of which we are aware—the words and mental images—are never sufficient for the function attributed to them. In every case, more is going on, as it were, 'offstage.' And what more is going on is, at least in certain cases, patently not expressible as a sentence or a picture." (PYLYSHYN, 2003, p. 436).

In *The language of thought* Fodor presents another reason to suppose that there is a language of thought different from any natural language. According to him,

> one cannot learn a language unless one has a language. In particular, one cannot learn a first language unless one already has a system capable of representing the predicates in that language *and their extensions*. And, on pain of circularity, that system cannot be the language that is being learned. But first languages *are* learned. Hence, at least some cognitive operations are carried out in languages other than natural languages. (FODOR, 1975, p. 64).

[139] In fact, Pylyshyn often extracts a stronger conclusion from his arguments, which is that "if the sentences do not express all that I know or intend, then what I know or intend must take some other form—a form of which I have no conscious awareness." (2003, p. 430). I think, though, that this stronger conclusion, that the vehicle of thoughts is not accessible to consciousness, is not well supported by the arguments. I assume that it is possible to have conscious thoughts that occur neither in images nor in natural language sentences.

So it is not only that we need to assume the existence of a language of thought because natural languages are too ambiguous for that role, or because we can think about things for which we have no words. According to Fodor, only by assuming the existence of a language of thought can we explain our ability to learn a first language.[140] Roughly, in his view, learning a first language involves, among other things, formulating hypotheses about the extension of the predicates of that language. So in order to learn what a predicate means, one already needs to be able to represent the extension of that predicate; one already needs to have a concept which is coextensive with the predicate: "learning (the word) 'green' requires having the concept GREEN" (FODOR, 2008, p. 137). I can only learn what "green" means if I can think about green. Pinker uses the same argument for the assumption that the language of thought must be different from a natural language. According to him, "if babies did not have a mentalese to translate to and from English, it is not clear how learning English could take place, or even what learning English would mean" (PINKER, 1994, p. 82). So the idea that emerges is that (at least some) concepts are prior to words, and that words are merely expressions of concepts. In learning a language, we learn, among other things, how to map a predicate to a concept.[141]

Pinker adds furthers problems for the assumption that natural languages are the medium of thoughts. According to him, "if thoughts depended on words, how could a new word ever be coined? How could a child learn a word to begin with? How could translation from one language to another be possible?" (PINKER, 1994, pp. 58). In addition, he draws attention to the fact that everyone has experienced the feeling of wanting to say something, but not being able to express in words what one wants to say. According to Pinker, this suggests that there is something, which is what one wants to say, that is different from what one actually says in words. For all these reasons, Pinker says that "English (or any other language people speak) is hopelessly unsuited to serve as our internal medium of computation." (PINKER, 1994, p. 78).

---

[140] Similarly, in a more recent book, he says: "here's another standard argument against 'language before thought': Language learning—including, in particular, first-language learning—takes a lot of thinking on the part of the learner. So, if you have to be able to talk before you are able to think, it follows that you can't learn a first language. This seems to be an embarrassment since, in point of fact, many children do so." (FODOR; PYLYSHYN, 2015, p. 14).

[141] Against the charge that this sort of explanation of meaning leads to an infinite regress, Fodor says "My view is that you can't learn a language unless you already *know* one. It isn't that you can't learn a language unless you already *learned* one." (FODOR, 1975, p. 65). The language of thought is where we stop.

The tip-of-the-tongue phenomenon is a manifestation of this feeling Pinker talks about, in which one has a clear sense of knowing what one is thinking about, in spite of being unable to retrieve the word that is used to express it. In its pathological manifestation, that is what happens to people who suffer from anomia, which is a form of aphasia that affects one's ability to recall words. Ashcraft (1993), a psychologist who suffered from a transient anomia caused by brain-related abnormalities, reports the subjective experience that accompanied it. He says that "the subjective experience consisted of knowing with complete certainty the idea or concept that I was trying to express and being completely unable to find and utter the word that expressed the idea or concept." (ASHCRAFT, 1993, p. 49). In spite of his inability to retrieve certain words, he reports how he noticed that something was wrong with him, and how he tried to test if he was alright by leaving his office and walking to the restroom. He also managed to recall his wife's phone number and call her, though he couldn't communicate with her properly. He claims that he managed to do all this without saying words to himself in inner speech. According to him, "attention/awareness, short-term memory and problem solving processes requiring a sustained 'train of thought' were relatively unaffected by the seizure." (1993, p. 54).

To sum up, the arguments presented in this section suggest that natural language sentences are not the vehicle of thought.

4. Problems with Carruthers' views

As I mentioned earlier, Carruthers advocates a cognitive conception of language, according to which natural language is the vehicle of conscious propositional thinking. He does accept that there can be a language of thought which is the vehicle of some of our thoughts, but he thinks introspection shows us that conscious thinking occurs in a natural language. As he puts it, "introspection informs us, in fact, that many of our thoughts are expressed entirely in natural language" (CARRUTHERS, 1996, p. 50), and that if conscious thinking takes place in English, then it doesn't take place in Mentalese (idem, p. 55), which is another name for the language of thought. But for the reasons I just presented, natural language sentences don't seem to exhaust all there is to thinking. Even if it is true that most people have the experience of inner speech, introspection itself doesn't show us that we don't need anything else in order to have conscious

thoughts, or that Mentalese is not involved in generating the phonological representations of English sentences in inner speech.

In reply to the observation that natural language sentences are not sufficient for conscious thought, Carruthers could then claim that sentences in a natural language, even if not sufficient, are necessary for conscious thoughts. In fact, this is what he claims:

> no one should want to claim that a tokened natural language sentence *is* (or is sufficient for) a thought. (Consider a monolingual speaker of Russian uttering a sentence of English, for example.) Indeed, defenders of the cognitive conception, too, should accept that the content of an inner tokened sentence will depend upon a host of more basic connections and sub-personal processes. Rather, the claim is that the sentence is a *necessary component* of the thought, and that (certain types of) reasoning necessarily involve such sentences. (CARRUTHERS; BOUCHER, 1998, p. 14).

Again, Carruthers' main evidence for this view is introspection: the fact that we often experience our own thought as couched in inner speech. But introspection doesn't show us that, were the phonological representations of natural language sentences gone, conscious thought wouldn't occur – it doesn't show us that natural language sentences are necessary for conscious propositional thought. And if we are going to rely on introspection, we should take anomia and tip-of-the-tongue cases seriously, which suggest that people can have conscious thoughts that don't involve natural language sentences. It is open to investigation whether all thoughts could occur consciously without natural language sentences, but it seems that at least some can. Assuming anomia reports are accurate, they suggest that natural language is not the vehicle of all thoughts, and also, against Carruthers, that it isn't even necessary for the occurrence of all conscious propositional thoughts (though it may be for some). It seems possible to have a concept consciously activated, while being unable to retrieve its phonological expression.

I don't want to get into the topic of consciousness here, but I would add that we certainly don't seem to be aware of ourselves talking in inner speech whenever we make decisions, or whenever we believe or desire something, and the absence of inner speech or visual imagery doesn't seem to make those decisions, beliefs and desires any less conscious. It's probably true that we often experience our own thoughts in a natural language, but, if we are to follow Carruthers and rely on introspection, we should conclude that it is probably false that we *only* experience our propositional thoughts in sentences in a natural language. Besides, Carruthers' view that

conscious propositional thought requires natural language sentences as its vehicle also has the counterintuitive consequence that animals are incapable of conscious propositional thoughts, and that infants only become conscious of certain things when they learn a language.

Carruthers also seems to be thinking that the common sense idea is that natural language sentences are necessarily involved in our conscious propositional thinking, and that in denying that, Fodor will need to explain "the systematic illusion that our conscious thoughts involve natural language" (CARRUTHERS, 1996, p. 61). But it is far from obvious that the common sense view is that strong. Again, it is not clear that introspection itself strongly commits us to the view that public language is necessary for conscious propositional thinking – even if we accept that inner speech is a common phenomenon. As I see the issue, accepting that LOT is the vehicle of thought doesn't force us to say that we are under a "systematic illusion" in our conscious thinking, that needs explaining. I can have inner speech in English and still accept that LOT is the real substance behind it; that LOT is what is causing the phonological representations of English sentences, and giving an interpretation and structure to them. So there is no need to deny that some of us at least sometimes have the experience of inner speech in a natural language. But we have to be careful about what this represents for the common sense view on thinking, and about what we can conclude from this about the vehicle of thought.

Carruthers also criticizes some of the arguments for the idea that natural languages are not the vehicle of thought, such as the argument from the ambiguity in natural languages. According to him, natural language sentences that occur in inner speech are not ambiguous:

> an imaged sentence is not, in any case, a merely phonological representation; rather, the content is represented in the image too. The imaged sentence, 'I shall go to the bank', carries with it one or other possible interpretation, just as the heard sentence would do. When we hear people speak, we do not just hear phonology, and then have to figure out an interpretation. Rather, what we hear is already interpreted. If someone says to me, 'I shall go to the bank', then I will immediately *hear* this as meaning one thing or another (depending on the context) - though I may have to think again if the meaning I hear is sufficiently incongruous. So it is, too, with inner speech. Ambiguous sentences always carry with them some particular interpretation when imaged. (…) the background dispositions which constitute my meaning the sentence in one way or the other may be realised in some sort of representational structure (of Chomsky's logical form, or LF, as it might be) which is already attached to the phonological representation. So the image is actually an interpreted object – it is not ambiguous. (CARRUTHERS, 1996, p. 58).

I think that, in some respects, there is no real disagreement here between Carruthers and the proponents of LOT, but rather a different view on what is to count as a public, or natural language sentence. Carruthers thinks that a natural language sentence is individuated by the sum of its phonological, semantic and syntactic forms. So sentences in English are not ambiguous when they occur in inner speech because they are in part constituted by their interpretation. He can then say that we think in English and that our thought is not ambiguous because he is taking the semantic level of analysis, or the interpretation of the sentence, to be part of what constitutes the sentence.

Proponents of LOT, on the other hand, are taking a natural language sentence to be a phonological form together with a surface syntactic form. The deep structure, or logical form, and meaning are given by LOT. This is why, for Fodor, some sentences are ambiguous, and unsuitable for being the vehicles of thought: their phonological and surface syntactic form can be given more than one interpretation. The ambiguity is in the surface form of the sentence. What disambiguates public sentences is the language of thought. For some, like Fodor and Pinker, the logical form of the sentence is actually a level of description that belongs to the thought behind the English sentence, and not to the English sentence itself. This is why English sentences can be ambiguous, whereas thoughts cannot.[142] So Carruthers and Fodor agree that thought is not ambiguous. But since Carruthers thinks that sentences are also not ambiguous, he sees no problem in saying that we think in English, whereas Fodor thinks English sentences are sometimes ambiguous, and therefore not suitable for being the vehicle of thought.

I think, though, that Carruthers is not taking as seriously as he should the asymmetry between a sentence that occurs in inner speech and a sentence that we hear in linguistic communication. While a sentence that occurs in inner speech is never subject to ambiguity, a heard sentence can be – even if infrequently, because contextual and background information are usually enough to disambiguate sentences. A sentence in inner speech comes attached to what it means simply because it is directly expressing concepts and LOT sentences. But the same doesn't necessarily occur with a sentence we hear. The fact that we can be unsure about the meaning of a heard sentence suggests that the phonological and surface syntactic form can be detached from a

[142] In *LOT 2* Fodor says in a footnote: "I think that LF [logical form] is a level of description not of English, but of Mentalese" (2008, p. 78, n. 50). In *Hume variations*, "it seems to be open whether there are ambiguous sentences, but it's closed whether there are ambiguous thoughts. So far as ambiguity is concerned, thoughts are where the buck stops." (2003a, p. 156).

particular interpretation. We can explain our misunderstanding of a sentence by our failure in attaching to it the meaning that its speaker intended it to have. A proponent of LOT would say, then, that it is the job of LOT to give an interpretation to the sentence. If the English sentence "I need to call Sarah" occurs to me in inner speech, I will have a specific Sarah in mind, even though I know many Sarahs. I don't need to clarify my thought, as I often need to clarify my request to my husband, when I ask him to call Sarah. Here, misunderstandings can (and often do) occur. And a reasonable explanation for this is that "Sarah" is a name that I use to express the concept I have of a particular person. And there happens to be a many-to-one relation between concepts of people I know, and the name "Sarah". In communicating my thoughts to my husband, he will face the problem, even if unconsciously, of attaching a concept to the word "Sarah", and the wrong matching can occur. The natural way to describe the situation here is to say that the word is ambiguous. If words were individuated also by what they mean, then they would never be ambiguous: "Sarah" meaning my mother, and "Sarah" meaning my friend from school would simply be two different words.[143]

Another reason to assume a detachment between words and concepts, or between the phonological level and the semantic level, is suggested by some recent studies. Yang et al. (2017), for instance, mapped the brain activity in English speakers when they read some sentences in their language and then fed the data to a computational model. They then recorded the brain activity from Portuguese speakers when they read sentences in Portuguese, and presented the model with their brain activity data. The model was able to say, from the Portuguese speakers brain activation data (which it hadn't encountered before), which sentence they were reading. This suggests that the brain areas responsible for the comprehension of a sentence in Portuguese were similar enough to the ones responsible for the comprehension of sentences in English, despite the obvious differences in phonological and surface forms. I think this can be taken as evidence that there is a common level of conceptual or semantic representations across speakers of different languages. This favors the view that takes words and sentences in a natural language to be individuated by their phonological and surface syntactic forms, while the semantic level and the level of logical form are taken to be levels of thought properly. In so individuating words and concepts, we can say that thoughts are entities that are sharable, even among individuals that

[143] There is an intuitive sense in which two people can share the same name, or in which "bank" and "bank" are the same word, what is different is what they refer to. The view that words are individuated by their phonological properties is, I think, more compatible with the common sense way of individuating words.

don't share the same natural language. These common semantic representations can then be mapped to phonological or written expressions in the particular language of each individual.[144]

We can, if we want, accept Carruthers' way of individuating words and sentences. But it certainly seems that thought can never be ambiguous while sentences at least sometimes appear to be – so he would have to explain why that seems to be the case. It also seems intuitive to suppose that I can, for instance, think about a specific Sarah without saying "Sarah" to myself, the phonological representation being perfectly dispensable. We are not always saying words to ourselves. And the evidence from tip-of-the-tongue and anomia also suggests that concepts can be detached from the words expressing them. Even if a word is partly constituted by the concept it expresses, some concepts at least can occur in conscious thoughts without being accompanied by words or sentences in a natural language – which suggests that Carruthers' view about what constitutes conscious propositional thought is not correct.

To sum up, language is probably not necessary for all thought, though it probably is, as a matter of fact, required for the acquisition of certain concepts. In addition, the arguments from the proponents of LOT reviewed so far, taken together, make a good case against the idea that natural language is the vehicle of all thought.[145] These arguments don't prove once and for all that all thoughts can occur to us without natural language as part of their vehicles. It could still be the case that some thoughts perhaps only play a role in our thinking when they are available to us in inner speech. It could also be the case that some unconscious thoughts involve, in addition to representations of logical form, representations of natural language sentences. Also, even if all

---

[144] One interesting question is whether the semantic representations that we use in language production and comprehension are the same as (or a subset of, or different from) the conceptual representations that we use in thinking, construed in a broader sense. That is, can semantic representations, that give meanings to words, be the vehicle of other cognitive processes that do not involve language? Were semantic representations used only in language production and comprehension, it might be more natural to count them, with Carruthers, as part of what individuates sentences and words in natural language. In perception and categorization, for instance, do we activate the same concepts that we activate in thinking? These are questions that I hope to investigate in future research.

[145] Or at least they make a good case against the idea that phonological representations are the vehicle of thought. The arguments don't necessarily oppose the view that we at least sometimes use natural language sentences to think, if we assume that words are partially individuated by their interpretations. We could say that we think in English when we experience inner speech in English sentences, as long as we take English sentences to be the sum of phonological, syntactic and semantic representations. But I'm assuming that it is more natural to separate words from concepts, and take words to be primarily phonological expressions of concepts. For one thing, as the proponents of LOT point out, there are more concepts than there are words. And we frequently coin new words, or name new objects and properties. In doing so, it seems reasonable to assume that before the naming event, there was already a concept, the word functioning simply as a label. But also, as I argued, it seems that semantic representations can occur in thinking without the phonological representation to which it is usually associated. If we accept this, we could say, about inner speech, that it is Mentalese plus a natural language.

thoughts that occur to us in inner speech could equally well have occurred without the phonological garment, so to speak, it could still be the case that inner speech has a relevant cognitive function, like helping memory. So the previous considerations don't refute the idea that inner speech is part of people's conscious lives, nor that it has a cognitive function. They only show that natural language sentences, when individuated by their phonological properties and surface form, are not sufficient to explain thought processes – so even when we have inner speech, there is more going on than mere sequences of words – and that we can perfectly well have some thoughts that don't have natural language as their vehicle, even conscious ones.

But even if we think Carruthers is wrong in saying that language has the cognitive role of being the vehicle of all conscious propositional thinking, that doesn't mean that the only function that's left for language is that of communicating thoughts. I think Carruthers' opposition between cognitive conception and communicative conception, with the former but not the latter accepting that language is the vehicle of thought, though useful, is in fact too crude. It doesn't exhaust the different positions in the logical space with regard to the relationship between thought and language. Carruthers thinks that "only if language is constitutively involved in the processes of thought itself will the cognitive conception of language be vindicated over the competing communicative picture." (1996, p. 18). Against Carruthers, one could support the view that language has a cognitive function, while denying that there are any thoughts that constitutively involve language. There are, for instance, different views as to whether and how the particular natural language one speaks affects non-linguistic cognitive domains. Some studies claim to show that speaking language L instead of language M makes one more prone, in some circumstances, of having cognitive feature X and not cognitive feature Y.[146] If something like this is in fact the case, we could say that the language we speak affects one or other cognitive domain without being necessarily involved in any thought. Whether or not it would be right to describe this kind of role of natural language as a cognitive role, the fact is that this is a kind of influence that the particular language one speaks perhaps has on thought that is not clearly contemplated by Carruthers' opposition between cognitive and communicative conceptions of language.

One question that I wish to pursue in the rest of the chapter concerns the degree to which the specific natural language we speak influences our thinking, or shapes some of our concepts. As

[146] For comprehensive reviews about the research in linguistic relativity, see for instance Wolff and Holmes (2011) and Everett (2013). Some classic studies in linguistic relativity include, for instance, comparing the performance in the sorting of objects by shape or substance, between speakers of languages that use classifiers and speakers of languages that don't. In the following sections we are going to explore the case of colors.

I'm going to suggest, one can accept the arguments presented by Fodor, Pinker and Pylyshyn, but at the same time accept that the language we speak shapes some of our concepts – though not in any strong, deterministic sense. I think the proponents of LOT give us good arguments against the view that natural language is the vehicle of all thought. It is true that language is ambiguous, that we can think about things for which we have no words, that it would be difficult to explain first language acquisition if we were completely incapable of conceptualization and thought prior to the acquisition of language. It is also true that languageless deaf people, aphasics and pre-linguistic children, as well as animals, exhibit all sorts of intelligent behavior, which can only be explained under the assumption that they are capable of at least some cognitive processes that are independent of language. But all that shows is that language and thought are not strictly the same thing – that language is not necessary for, nor involved in, all thoughts. It doesn't show that language cannot influence some thoughts or concepts that we have.

One possibility that is still open, even with the acceptance that we think in a language of thought, is that natural languages might be a shaper of at least some concepts. As we've seen, Whorf formulates a view according to which language has the function of shaping our ideas in a strong sense, for without it there would be no concepts by means of which we classify things in the world. Language cannot be doing all this work, but it may be playing a similar role for at least some of our concepts (the example I'm going to explore from section 8 on is that of color concepts). The language we speak may bias us to categorize certain properties in a certain way, because, in speaking the language we do, we get accustomed to categorizing things in the way our language requires us to. It can also force us to pay attention to certain features of the world that maybe we wouldn't attend to if we spoke a different language, or if we spoke no language at all. So in learning certain words, we learn to attend to certain features of the world that otherwise we might have completely ignored. But we also come to group certain things together that don't necessarily belong together. Again, this is not to say that we would have no concepts or categories at all if it weren't for the natural language we learned, nor that our native language is involved in, or is the vehicle of, the concepts that it has the power to shape. This is also not to say that we are incapable of conceiving anything that is not admitted by our language – if that were the case, as Pinker pointed out, people would not be able to learn a foreign language. It is an empirical question the degree to which the particular language we speak can influence the way we think.

But, as we will see, there is already some evidence that that influence exists, albeit not in any strong deterministic sense.

The kind of influence that I'm suggesting language has on thought is weaker than the cognitive function that Carruthers attributes to natural language – since it is possible for the language we speak to shape some of our concepts without being the vehicle of the thoughts involving those concepts. So natural language can play the role of shaping our concepts, in the sense that some concepts inherit their extensions from the extensions of the words in a language. This can perhaps be characterized as a kind of cognitive function of language, but it is one that doesn't imply that natural language is the vehicle of the concepts that it shapes. That is, one can accept that natural language is used to shape some of our concepts while also accepting that these concepts are symbols in a language of thought, that are expressed or communicated by words in a natural language. But while this view is weaker than Carruthers' cognitive conception of language, it is stronger than the merely communicative conception that Fodor seems to hold. As I read him, Fodor leaves very little room for any influence of natural language on thought.

What I will do in the next sections is to review some of Fodor's views about concepts, and the relation between the semantics of thought and the semantics of language, in order to assess them in light of some empirical findings. I'm going to argue that, at least in the case of color concepts, language is not simply giving us a label for a concept we already had, but it is modifying our concepts and making them concepts that we share with other members of our linguistic community.

5. Fodor on innateness, complex and primitive concepts

We have seen so far that Fodor and others hold the language of thought hypothesis, according to which thought occurs in a system of mental representations that have a syntactic constituent structure and a compositional semantics. In the previous sections, we briefly saw some of the possible views on the relationship between thought and language, as well as the arguments given by Fodor, Pylyshyn and Pinker for the idea that the language of thought is not a natural language. In the rest of this chapter, I'll explore some of Fodor's views about words and concepts, as well as his views on innateness, and on the semantics of thought and the semantics of language.

I will then criticize some of these views in light of some empirical observations about color words and concepts.

As we've seen, one of Fodor's arguments for LOT is that we need to assume the existence of a language of thought in order to explain the acquisition of a first natural language. Fodor argues that we already need to have a language (the language of thought) in order to be able to learn a natural language. Fodor basically thinks that, if we are to be able to learn a predicate X in a given language, we already need to have a concept, coextensive with that predicate. Concepts are mental entities, symbols or representations in the language of thought, which are sharable.[147] According to Fodor, whatever can be expressed by a natural language can also be expressed by the system of mental representations. So Fodor doesn't think that learning a language can increase the expressive power of our system of mental representations.

We could then say, against this view, that it doesn't seem very likely that there is no cognitive advantage in learning a natural language. As we've seen in previous sections, it seems that natural language is at least a prerequisite for us to have certain thoughts, thoughts that we wouldn't, as a matter fact, have, had we not learned a natural language. Trivial examples of such thoughts would be thoughts about language, for instance the thought that English has over two thousand words. Other examples are thoughts about scientific entities, like the thought that all living things have DNA. If Fodor's view is right, according to which one already needs to have concepts in order to learn what a predicate means, it is unclear how learning a natural language could give us access to concepts that we didn't have before.[148] But Fodor doesn't claim to be committed to a view according to which learning a natural language gives us no cognitive advantage. On the contrary, in *The language of thought* he concedes that learning a natural language gives us some cognitive advantage over an inarticulate organism. A possible advantage, according to him, is that certain words can presumably serve to abbreviate complicated formulae of the language of thought, which our limited computing memory wouldn't possibly be able to process were they in their original complicated format. As Fodor puts it,

---

[147] There is no need to infer, from the assumption that concepts are mental tokens, that they refer to private objects in a private experience, inaccessible to anyone else. Someone who accepts that there is a language of thought is not thereby committed to solipsism or skepticism about the existence of other minds or the external world. This is why I think Wittgenstein's objections to a private language don't apply here. I am here assuming, with Fodor, that concepts are sharable entities. When I think about apples, and Mary thinks about apples, we are tokening concepts of the same type, whose referent is a property in the world, and not something only accessible to the person who has the thought. To the extent that attributing thoughts and concepts to other people usually allows us to explain and predict their behavior successfully, the assumption that people share thoughts and concepts is at least reasonable.

[148] I will insist on this point in section 7.

> It is no news that single items in the vocabulary of a natural language may encode concepts of extreme sophistication and complexity. If terms of the natural language can become incorporated into the computational system by something like a process of abbreviatory definition, then it is quite conceivable that learning a natural language may increase the complexity of the thoughts that we can think. (FODOR, 1975, p. 85).

Here it seems that Fodor is even acknowledging that natural language can be the vehicle of some thoughts. Some words can be used perhaps to abbreviate complex concepts, and using words to think can allow us to think thoughts that in practice, though not in principle, we wouldn't be capable to think. Fodor even acknowledges, maybe surprisingly, that the natural language we learn may determine the kinds of concepts that we have, which is an idea close to linguistic relativism:

> there [isn't] any need to deny the Whorfian point that the kinds of concepts one has may be profoundly determined by the character of the natural language that one speaks. Just as it is necessary to distinguish the concepts that can be expressed in the internal code from the concepts that can be entertained by a memory-restricted system that computes with the code, so, too, it is necessary to distinguish the concepts that *can* be entertained (*salve* the memory) from the ones that actually get employed. This latter class is obviously sensitive to the particular experiences of the code user, and there is no principled reason why the experiences involved in learning a natural language should not have a specially deep effect in determining how the resources of the inner language are exploited. (FODOR, 1975, p. 85).

But even though it may seem like Fodor is conceding a lot to linguistic relativism, in light of other things he says in the book, and elsewhere, I think he should not be interpreted as saying that natural languages have any influence on the primitive concepts we have. As it is well known, at this point Fodor thought that primitive concepts could not be learned, so they had to be innate. Roughly, he conceived concept learning as a rational process that occurs through hypothesis formation and confirmation. The idea, then, was that in order to form a hypothesis about whether something is or is not in the extension of a primitive concept, the concept would already have to be represented in the hypothesis. So concept learning couldn't be a means by which primitive concepts are acquired, because it presupposes the availability of the concept that is

supposedly being learned. Fodor took this to mean that primitive concepts could not really be learned, so they had to be innate.

Also, Fodor thought that only the assumption that at least some concepts are innate could explain our ability to learn a natural language – assuming that learning a language involves formulating rules about the extensions of its predicates, which in turn requires one to be able to represent those extensions. Unlike most proponents of the linguistic relativity hypothesis, Fodor thinks that infants already have, prior to the acquisition of a first language, "a complexly structured and endogenously determined representational system." (1975, p. 86). This is supposed to explain, among other things, the assumed existence of linguistic universals. So, given that for Fodor primitive concepts were assumed to be innate, and prior to the acquisition of a first language, he cannot be interpreted as saying that the particular natural language we learn can influence the repertoire of primitive concepts we have.

My best guess, then, is that when Fodor seems to be conceding that the natural language we learn can influence the kinds of concepts we have, he was thinking, in *The language of thought*, that it can affect our *complex* concepts. That is, Fodor seems to be assuming that languages don't differ much in their lexical items that express primitive concepts. At the same time, he concedes that there are lexical items that can be defined from simpler ones. He also thinks that the same "process of abbreviatory definition" that occurs in natural language can occur in our concepts. So the distinction between primitive and complex concepts would correspond to the distinction between undefined and definable concepts. So maybe learning a language can be a way of triggering one to combine primitive concepts to form new, complex concepts, through a process of definition. Fodor says, for instance, that he didn't want to be committed to the view that children are born with concepts such as AIRPLANE. Assuming "airplane" can be defined as "flying machine", then perhaps the concept AIRPLANE can be introduced by definition from the concepts FLYING and MACHINE, in which case people would only need to be born with the primitive concepts into which the complex concept AIRPLANE decomposes. So assuming that there are primitive concepts, out of which complex concepts could be defined, the complex concepts would not be required to be innate. Complex concepts could then be the concepts that can be influenced by the language one speaks. Perhaps acquiring a natural language could be a way of making it practically possible for us to think certain thoughts, by means of what Fodor calls abbreviatory definitions, which otherwise would be too complex for us to entertain. But still,

the impact of the particular natural language we learn would be limited to our complex concepts, since the primitive concepts, which are expressed by the undefined lexical items, are supposed to be innate, and therefore not subject to the influence of a particular language.[149]

It's apparent, from the discussion so far, that the language of thought hypothesis rests on the assumption that there is a finite number of atomic or primitive concepts, out of which an unlimited number of complex concepts can be formed. The natural question is what distinguishes primitive symbols from complex ones. In Fodor's view, primitive concepts are concepts that have no internal structure, or that cannot be defined in terms of others. Complex concepts, on the contrary, are the ones that have internal structure, or that can be defined in terms of others. It is useful to introduce, at this point, the distinction between lexical concepts and phrasal concepts. Lexical concepts are the concepts that are expressed by monomorphemic expressions of a natural language (e.g. the concept COW is expressed by "cow" in English), whereas phrasal concepts are the ones expressed by morphologically complex expressions, or multimorphemic predicates in the same language (e.g. BROWN COW is expressed by "brown cow"). Most phrasal concepts seem to be complex, because their semantic properties can be derived from the lexical concepts that constitute them.[150] The question then is whether all (or most, or no) lexical concepts are primitive, or whether all (or most, or no) lexical concepts are in fact structured around, or constructed out of simpler concepts, despite the apparent simplicity of their linguistic expressions.[151]

---

[149] It is interesting that Fodor argues that, even if we accept that words and concepts have definitions, it does not follow that their meanings are psychologically represented by their definitions. For example, when I hear the sentence "John is a bachelor", according to Fodor, it's not reasonable to assume that the message of this sentence is internally represented as JOHN IS AN UNMARRIED MAN (assuming the definition of "bachelor" is "unmarried man"). If definitions were represented at a real psychological level when we hear a sentence, we should assume that it would be more difficult (or that it would take more time) to understand, say, the sentence "John is a bachelor" than the sentence "John is unmarried", because the first, when internally represented, would be more complex, having UNMARRIED as one of its constituents. But that doesn't seem to be the case (cf. FODOR, 1975, p. 147). So Fodor, in *The language of thought*, accepts that there can be a definitional or abbreviatory process occurring in thought as well as in language, allowing new concepts to be introduced to the internal vocabulary. But once these complex concepts are introduced, they would end up functioning as unstructured symbols, with no need for the definiens to be tokened whenever the definiendum is. Complex concepts would be connected to their definitions by what Fodor calls "meaning postulates", which are inference rules. In the end, according to him, "the primitive vocabulary of the internal representational system is comparable in richness to the surface vocabulary of a natural language" (FODOR, 1975, p. 133). If this is true, according to Fodor, "it goes some way toward explaining why natural languages are so easy to learn and why sentences are so easy to understand: The languages we are able to learn are not so very different from the language we innately know, and the sentences we are able to understand are not so very different from the formulae which internally represent them." (FODOR, 1975, p. 156).

[150] Though this doesn't apply to idioms. It also doesn't seem to be the case that all compound words are definable from their compounding words, like "watermelon", "firefly", "Walkman" or "notebook".

[151] Fodor concedes that, in some cases, concepts that are expressed by phrasal expressions in English can by expressed by monomorphemic predicates in language X, and vice versa. But as he says, the distinction between lexical and phrasal concepts is mainly heuristic. What makes a concept primitive or complex is not how it is

In *The Language of Thought*, Fodor was assuming that not all lexical concepts are primitive, but rather that some at least are in fact complex, because they can be decomposed into simpler ones, that define them (he thought, as I mentioned, that AIRPLANE presumably could be an example of a complex concept, which could be decomposed into the concepts FLYING MACHINE). So here only primitive concepts were presumed to be innate, but it was a possibility that a good number of lexical concepts would turn out to be complex (because definable), and therefore not innate. It was still possible that learning a natural language could influence a considerable number of the lexical concepts we have.

The situation becomes even more hopeless against the possibility of an influence of natural languages on the concepts we have in his article "The present status of the innateness controversy" (1981). Here Fodor comes to think that there aren't many definitions around. It's often easy to find counterexamples to necessary and sufficient conditions proposed for the application of a given term. And if there aren't that many definitions, the number of primitive, unstructured concepts has to be bigger than he once thought. The idea now is that most – if not all – lexical items (or monomorphemic words) express primitive concepts, which have no internal structure because they are not definable. So Fodor became skeptical of the idea that we are going to find many lexical concepts that can be decomposed into more primitive ones that define them. He came to think that there is a more or less reliable correspondence between primitive concepts and words; most lexical concepts are primitive, whereas before, the possibility was open that a considerable number of concepts expressed by monomorphemic words could still be complex, if they were definable.[152] Since, for him, primitive concepts cannot be learned, and lexical concepts are supposed to be primitive, since they are not defined and have no internal structure, Fodor is led to hold an extreme nativism, according to which probably most lexical concepts are innate. As he notes,

> if a concept belongs to the primitive basis from which complex mental representations are constructed, it must *ipso facto* be *un*learned. (To be sure, some

---

expressed in language, which can vary from language to language, but whether the concept in question can be defined, or decomposed into simpler units. If it cannot be defined, then it is primitive; if it can, then it is complex.

[152] They could be complex, but not in the sense that definable lexical items were internally represented by structured primitive concepts. He already thought in *The language of thought* that there was a correspondence between the lexical items of a natural language and the symbols in the language of thought. But he thought that concepts could be introduced into the internal vocabulary by definition from primitive ones. He called definable concepts "complex", even though they didn't have to have as constituents, whenever they were tokened, the concepts out of which they were defined.

> versions of RTM are rather less up front in holding this than others.) Prima facie, then, where a theory of concepts draws the distinction between what's primitive and what's not is also where it draws the distinction between what's innate and what's not. (FODOR, 1998a, pp. 27-8).

If most lexical concepts are primitive, and therefore innate, then a wide range of concepts, basically most concepts that are in English expressed by lexical items, are not going to be subject to the influence of the particular natural language one learns.[153]

One could argue against Fodor's identification between unstructured and undefined concepts, or structured and definable. It could be said that a concept can have an internal structure, while still being undefinable. This is the view implicit in the prototype theory of concepts, according to which concepts are structured not around their definitions, but around their best examples, or prototypes. Fodor also raises some problems against this view. What he says is basically that we need to assume that the meanings of complex concepts are inherited from the meanings of their constituents, that is, we need to assume compositionality, in order to explain the productivity and systematicity of thought. If concepts were prototypes, the prototype of a complex concept would have to be inherited from the prototypes of its constituents. But, as Fodor notes, complex concepts typically don't have prototypes (e.g. MOOD ON TUESDAYS AFTER LUNCH), and, if they do, in most cases the prototype of a complex concept is not inherited from the prototypes of its constituents. Fodor famously gives the example of the complex concept PET FISH. The prototypical pet (say, dog) and the prototypical fish (say, a salmon) don't really contribute to the extension of the prototypical pet fish (a goldfish). The prototypical pet fish has features that are not present in the prototypical pet or in the prototypical fish. Because prototypes don't compose, Fodor concludes that concepts are not prototypes.[154] Also, according to Fodor,

---

[153] We can interpret Fodor as saying that, in English, primitive concepts are typically expressed by monomorphemic words – though sometimes by idioms or multimorphemic and compound words. Perhaps other languages may make a heavier use of the latter forms of expression to express primitive concepts. What will ultimately be used to specify whether or not a concept is primitive is its undefinability, and not how it is expressed in a given language. It just happens that lexical items in English are typically undefined, which indicates that they are expressions of primitive concepts. That a concept is expressed by a monomorphemic word in English is simply a reliable indication that it is primitive, and therefore innate. A concept will not, of course, be innate in English speakers but acquired in Mandarin speakers, in case it is expressed by a monomorphemic word in English, but not in Mandarin. It would be interesting to explore some examples in detail. For instance, names for meals are compound in some languages but not in others: "dinner"/"Abendessen", "lunch"/"Mittagessen", "Breakfast"/"café da manhã". Are their respective concepts simple or complex? Fodor doesn't give us any general procedure for deciding which concepts are primitive.

[154] For an attempt to defend the prototype theory of concepts against the charge that prototypes don't compose, see Prinz (2012).

knowing what the prototype of a complex concept is is not essential to understanding a phrase that expresses a complex concept. Many concepts have prototypes, but that doesn't mean that concepts are prototypes, or are structured around prototypes.[155]

Also, it can be added that it is not even clear that all primitive concepts have prototypes. What is the stereotype of AIR, or CARBON, or DVD? And even assuming that they do have prototypes, prototypes are not always shared, so if concepts were prototypes, this would speak against the idea that people usually share beliefs and concepts. It seems that the owner of a fig farm and I can share the belief that fruit is good for you, even if her prototype of fruit is a fig, whereas mine is a banana. So even if concepts have prototypes, having a particular prototype is not something that is essential to the identity of a concept. It seems better to say that FRUIT and BANANA are different concepts. Their being often associated by me is what makes BANANA a prototypical fruit for me. But that doesn't mean that FRUIT has BANANA as one of its constituents; that the concept FRUIT contains in itself the concepts of particular fruits, with its center in the BANANA concept.[156] I could have the concept FRUIT even if I didn't have the concept BANANA.

Lexical concepts are then taken to be primitive concepts to the extent that they are not structured around their definitions, or around prototypes. And, as we've seen, Fodor thinks that primitive concepts are innate – because he argued that they couldn't be learned by hypothesis formation and confirmation that, according to him, is the only way concepts could conceivably be learned. But it is important to clarify that by saying that primitive concepts are innate, Fodor didn't mean to say that they were literally present at birth. What he meant was that mere experience with properties in the world is enough to trigger the acquisition of primitive concepts. He thought, though, that experience, being so varied, is not sufficient to explain the uniformity of the concepts we end up having. Given that experience itself doesn't seem sufficient to guarantee that concepts will be uniformly created across different people (which he is assuming is what happens), due to different individual histories, it seems that primitive concepts need to, in some sense, come from within. We come to the world predisposed to form the primitive concepts we do

---

[155] An alternative way conceiving prototypes is as associations. So robin is a prototypical bird because there is usually a strong association between the concept BIRD and the concept ROBIN.

[156] Also, the tokening of a belief doesn't seem to involve the tokening of the prototypes of each concept that constitutes that belief – but it seems that that should be the case, if concepts really were prototypes. That is, assuming all primitive concepts are prototypes, to token a concept would be to token its prototype. But even if apple is the prototypical fruit, that doesn't mean that I token the concept APPLE whenever I token the concept FRUIT, even if I may be more prone to token one after tokening the other.

when exposed to certain properties in the world. So Fodor is not denying that experience is involved in the acquisition of primitive concepts. These concepts are taken to be innate in the sense of being simply triggered by experience, by brute causal connections between innate biological mechanisms and certain properties in the world, as opposed to being learned by intentional rational processes.[157]

Color concepts, such as RED and GREEN, are among the typical examples of primitive concepts often given by Fodor (cf. 1998a and 2008, p. 137 and p. 139). Again, Fodor doesn't say that we are born with color concepts. He thinks we are born with biological mechanisms that lock us to the property of *redness*, say, once in contact with red things. Color concepts are, he thinks, simply triggered by the experience of colors. Given our innate mechanisms, experience with color properties is sufficient for us to acquire color concepts, with no learning required. According to Fodor,

> all that's required for us to get locked to *redness* is that red things should reliably seem to us as they do, in fact, reliably seem to the visually unimpaired. Correspondingly, all that needs to be innate for RED to be acquired is whatever the mechanisms are that determine that red things strike us as they do; which is to say that all that needs to be innate is the sensorium. (…) the 'innate sensorium' model suggests that the question how much is innate in concept acquisition can be quite generally dissociated from the question whether any *concepts* are innate. The sensorium is innate by assumption, and there would quite likely be no acquiring sensory concepts but that this is so. But, to repeat, the innateness of the sensorium isn't the innateness of anything that has intentional content. Since the sensorium isn't an idea, it is a fortiori not an *innate* idea. So, strictly speaking, the innate sensorium model of the acquisition of RED doesn't require that it, or any other concept, be innate. (FODOR, 1998a, p. 142).

I call attention to this particular example because, as I'm going to argue in the following sections, I think it is not true of color concepts that they are simply triggered by experience – though it is true of color perception: in order to see red, I only need to be exposed to red things (assuming I have normal color vision). As I'm going to argue, natural language is often part of what mediates my locking to, say, the property *greenness*, green things not being sufficient to trigger the concept GREEN in me, even given normal vision.

---

[157] According to Fodor, "if much of the conceptual repertoire is triggered, it's understandable that it should be invariant over wide-and relatively chaotic-variation in the individual histories of organisms belonging to the same species. To put it crudely: the concept isn't coming from the environment, it's coming from the organism. All the environment does is provide the triggers that release the information." (1981b, p. 280).

In any case, Fodor's views on innateness changed over the years. Whereas in *The language of thought* and "The present status" his view was that primitive concepts had to be innate, because they could not be learned, in *LOT 2* he maintains that they could be acquired. He still thinks concepts cannot be learned by hypothesis formation, and is actually firmer against it – now he thinks that complex concepts cannot be learned either, for the whole notion of concept learning is confused. But Fodor now thinks that learned and innate are not the only alternatives – according to him, concepts can be acquired by other means.[158] In *LOT 2*, he actually gives a positive account of how concepts could be acquired, without being learned and without being innate. According to him, perhaps we start by learning a concept's prototype, and learning a prototype is possibly a stage in the process of concept acquisition, the final stage being our locking to a property. As he puts it, "concept acquisition might proceed from stereotype formation to concept attainment. (…) acquiring a concept is, fundamentally, locking to the property that the concept expresses" (FODOR, 2008, p. 151). The locking part of the process of concept acquisition is supposed to be explained by a "not intentional (and hence, a fortiori, not inferential) neurological process" (FODOR, 2008, p. 151).

This view of concept acquisition is certainly more compatible with the idea that the language we speak can have some influence on the concepts we have. A consequence of the view that most lexical concepts are innate was that language could exert no influence on the acquisition of those concepts. In his new view, however, language could be, for instance, what enables us to pick the prototypes of certain properties in the world. But I'm hesitant in attributing to Fodor the view that natural language can be part of what determines the extension of at least some of our primitive concepts, because of his remarks against the possibility of an influence of the semantics of language on the semantics of thought, which I will discuss in the next section. Also, even if Fodor is no longer saying that a huge number of concepts is innate, I think he would presumably stick to the idea that color concepts are good examples of innate concepts, for he says that "even hard-core empiricists (like Hume) held that sensory concepts are innate" (FODOR,

[158] In *LOT 2*, he says "there are all sorts of mind/world interactions that can alter a conceptual repertoire. Concept learning (if there is such a thing) is one, but it's certainly not the only one. Other candidates conceivably include: sensory experience, motor feedback, diet, instruction, first-language acquisition, being hit on the head by a brick, contracting senile dementia, arriving at puberty, moving to California, learning physics, learning Sanskrit, and so forth indefinitely." (FODOR, 2008, pp. 131-2).

2008, p. 131). If that's the case, then the point I'm going to make against the idea that color concepts are innate is still an attack on his current views, and not simply his old ones.[159]

6. Semantics of thought and semantics of language

The topic of the semantics of thought has occupied Fodor in several of his writings and deserves a much more detailed exposition than the one I'm going to give it here. My aim here is to briefly highlight some of Fodor's main views on the semantics of thought and its relation to the semantics of language, in order to challenge them in the following sections.

We've seen that in *The Language of Thought* Fodor holds that concepts are prior to words, for he thinks this is part of what explains our ability to learn a first language. In *LOT 2* he continues to hold the same view. Learning a first language still involves, among other things, formulating hypotheses about the extension of the predicates in that language. So, again, in order to learn what a predicate means, one already needs to be able to represent the extension of that predicate; one already needs to have a concept which is coextensive with the predicate: "learning (the word) 'green' requires having the concept GREEN. This (…) strikes me as patently uncircular and true" (FODOR, 2008, p. 137).

Fodor also thinks that thoughts are more directly connected to the things they represent than natural language words and sentences are. Assuming that there are causal links that connect a representation to its referent, Fodor thinks these links are shorter in the case of mental representations than in the case of linguistic representations. In fact, according to him, linguistic representations only connect to their referents via mental representations:

> the chains that connect tokenings of mental representations to their semantically relevant causes are typically *shorter than* (indeed, are typically links in) the chains that connect tokenings of English sentences to their semantically relevant causes. This is the principal reason why it is mental representations, and not the formulas of any natural language, that are the natural candidates for being the primitive bearers of semantic properties. (FODOR, 1987, p. 100).

[159] In section 12, I will also consider the possibility that Fodor would allow some influence of the semantics of language on thought, allowing language to have an influence on the prototypes we learn. But I will argue that his positive account of concept acquisition doesn't seem very likely when applied to color concepts.

Fodor is now even more committed to the view that the semantics of thought is prior to the semantics of language, even suggesting that natural languages, strictly speaking, have no semantics. Words only have the semantic contents they have because of the concepts they express, and not the other way around.

> the Mentalese story is not just that the content of thought is prior to natural-language content *in order of explanation*; the Mentalese story is that the content of thought is *ontologically* prior to natural-language meaning. That is, you can tell the whole truth about what the content of a thought is *without saying anything whatever about natural-language meaning*, including whether there is any. (FODOR, 1998b, p. 68).

> the semantics of thought is prior to the semantics of language. So, for example, what an English sentence means is determined, pretty much exhaustively, by the content of the thought it is used to express. The corresponding implication is that semantics is essentially a theory of thoughts, the contents of which are, I suppose, *not* determined by norms and conventions. Quite possibly English has no semantics, some appearances to the contrary notwithstanding. (FODOR, 2008, pp. 198-9).

> forms of speech inherit their semantic contents from the concepts and thoughts that they express, not vice versa. (FODOR; PYLYSHYN, 2015, p. 12).

The idea is that most words express primitive concepts which are coextensive with and prior to them. To borrow Searle's terminology, thoughts and concepts have intrinsic intentionality, whereas words and sentences only have semantic properties in a derivative sense.[160] And if words and sentences derive their semantic properties from concepts and thoughts, then how could the semantics of language have any influence on thought? So even though, as we've

[160] As Searle puts it, "since sentences – the sounds that come out of one's mouth or the marks that one makes on paper – are, considered in one way, just objects in the world like any other objects, their capacity to represent is not intrinsic but is derived from the Intentionality of the mind. The Intentionality of mental states, on the other hand, is not derived from some more prior forms of Intentionality but is intrinsic to the states themselves." (1983, p. vii). This idea is hardly new. Ockham, for instance, says that "spoken words are signs subordinated to concepts or intentions of the soul (…). Spoken words are used to signify the very things that are signified by concepts of the mind, so that a concept primarily and naturally signifies something and a spoken words signifies the same thing secondarily." (1974, p. 50). This idea probably originates in Aristotle, who says that "spoken sounds are symbols of affections in the soul, and written marks symbols of spoken sounds. And just as written marks are not the same for all men, neither are spoken sounds. But what these are in the first place signs of – affections of the soul – are the same for all; and what these affections are likenesses of – actual things – are also the same." (1963, p. 43).

seen, Fodor says in *The Language of Thought* that the particular language we speak may determine the kinds of concepts we have, in *LOT 2* he says that

> There is no reason to suppose that 'how you think' or 'what you can think about' depends on what language you speak. Nothing but the semantics of Mentalese determines what one can think, or think about, and the semantics of Mentalese is prior to the semantics of English. (FODOR, 2008, p. 218).

Fodor attempts to provide, in several places (specially in *Psychosemantics* and *Theory of Content*), a semantic theory for concepts – since we cannot expect that their meanings will derive from the meanings of the natural language terms that express them. He holds an atomistic theory, according to which the meaning of a concept is independent from the meaning of any other concept, roughly because he thinks that if meanings were determined holistically, in relation to other concepts, it wouldn't be possible to preserve the folk psychological idea that people share concepts, as well as propositional attitudes. He also wants a semantic theory that preserves naturalism – that assumes that meaning is a property that can be treated scientifically, just like any other. Fodor is looking to provide at least sufficient conditions for a mental representation to mean what it does, and these conditions are to be presented in a non-intentional vocabulary, as naturalism requires. He comes to support a version of the causal theory of meaning, according to which a symbol X means *x* because *x* things cause tokenings of X. In order to deal with cases of misrepresentation, where a thing *y* can cause the tokening of X, Fodor introduces the idea that cases of misrepresentation are asymmetrically dependent upon cases of veridical representations (*y* things would not have caused tokenings of X if *x* things had not caused tokenings of X).

He has argued for a while now that reference is the only semantic property concepts have (there is no such thing as sense, or meaning, conceived as a semantic notion), and that reference is supposed to be reduced to causal relations between the mind and the world: "reference is a causal relation between referents-in-the-world and tokens of the symbols that refer to them" (FODOR; PYLYSHYN, 2015, p. 85). What a concept refers to, or represents, is essential for that concept's identity: "thoughts and concepts are individuated by their extensions *together with their vehicles*" (idem, p. 74). Reference is ultimately a causal relation between properties or individuals in the world and symbols in the language of thought, which holds regardless of what mechanisms mediate the relation. To the extent that words in a natural language also have semantic

properties, these semantic properties are inherited or derived from the semantic properties of the concepts that they express.

7. Some initial critical remarks

It is worth noting that Fodor's view on language learning makes it difficult to see how one can acquire new concepts through language. As I remarked in section 2, it was supposed to be a point of agreement between proponents of the communicative conception of language, such as Fodor, and proponents of the cognitive conception, that language is an important vehicle for making us capable of thinking about certain things. But if Fodor is right and I need to have a concept in order to learn a word that expresses that concept, it seems that words can never affect my conceptual repertoire. So it is difficult to see how this view can be compatible with the idea that language is a prerequisite for us to acquire certain concepts – even if language is not involved or is not constitutive of those concepts. If learning the word "atom" requires having the concept ATOM, if concepts are always prior to words, then it seems that language cannot really play any role in the acquisition of the concept ATOM.

Fodor's view on predicate learning is in fact very counterintuitive, for it seems quite obvious that we acquire new concepts all the time through linguistic communication. As we've seen, language really seems to be a prerequisite for us to acquire some concepts, like concepts about black holes, subatomic particles and presidential elections. That is specially the case for scientific and abstract concepts. In fact, most of the concepts we learn, even those that are about concrete entities, are not typically learned by direct causal contact to the properties and individuals that those concepts refer to. Even concepts that we could conceivably have acquired by non-linguistic means are often acquired through language. I can teach my 6 year-old niece that there is this cool animal called "echidna" that lives in Australia, which is one of the few mammals that lay eggs. I can tell her that it looks a little bit like a porcupine, and that it has no nipples, so babies get their milk from their mother's skin pores. My niece presumably ends the conversation with a new concept, one that she didn't have before, without ever having seen an echidna in her life – nor me, for that matter. She went from a stage of having no concept for echidna to a stage of having a concept for echidna as a result of a brief conversation. And this kind of concept acquisition happens all the time. If linguistic communication leads us to acquire

concepts we didn't have before, then it is not clear that, in order to learn a predicate, we already need to have a concept coextensive with that predicate.

What could Fodor say about this? In *LOT 2*, he makes a distinction between mediation and constitution, saying that "all sorts of things can mediate causal linkages between concepts and their referents without being in any sense constitutive of the identity of the concept" (2008, p. 142-3). So he could say that language is de facto *always* part of what mediates my acquisition of some concepts, such as ATOMS. In the case of other concepts, such as ECHIDNA, language just so happens to *sometimes*, though not always, be part of the causal chain that connects us to echidnas – some people at least have certainly acquired their ECHIDNA concept by direct contact with echidnas.[161] Someone like Fodor can then concede that language is responsible for our acquisition of perhaps most of the concepts we have, while denying that language is constitutive of what those concepts are. Words could be simply links that initially connect us to certain properties in the world, properties that we haven't yet experienced directly, being afterwards dispensable. That is compatible with a communicative conception of language.

But this reply wouldn't really address the point of the criticism. If we acquire concepts through linguistic communication, Fodor's view on what learning a predicate consists in cannot be true for all predicates. Fodor's conception of predicate learning may work, I think, for explaining how we learn predicates for certain properties we are surrounded by, and about which we can think before having learned any words. It explains how some of our first words can be learned: we start thinking about things that surround us, and then we name them. It also explains our coining of new words. But it clearly doesn't work for "atom", or for my niece's learning of "echidna". It is likely true that learning a word consists in attaching a concept to it, but that

[161] A child growing up in a place where echidnas are ubiquitous could conceivably acquire the concept ECHIDNA purely from interactions with echidnas, without ever having learned the word "echidna". So linguistic communication is not necessary for the acquisition of the concept ECHIDNA. The explanation of the acquisition of the concept ECHIDNA in those circumstances would fit Fodor's conception of innateness: experience with echidnas could be sufficient to trigger in the child the concept ECHIDNA. Should we say, then, that the concept ECHIDNA was innate in my niece, even if she acquired her ECHIDNA concept as a result of linguistic communication, because, were she exposed to the appropriate circumstances – an environment in which echidnas are ubiquitous – her concept ECHIDNA would have been simply triggered from causal encounters with echidnas? I think Fodor would probably say that the mechanisms for acquiring certain concepts are innate, but that that doesn't mean that the concepts themselves are innate. So sometimes concepts can be acquired from those mechanisms (when one is directly in causal contact with their instances), but sometimes they can be acquired indirectly, for instance through linguistic communication. But still, Fodor's view on concept acquisition doesn't seem appropriate to deal with concepts that are acquired by linguistic communication, as I'm going to indicate. Also, it seems that Fodor should be more specific about the nature of the innate mechanisms. Are the innate mechanisms that are responsible for my acquisition of the concept ECHIDNA, when in contact with echidnas, general mechanisms, or specific mechanism that apply only to animals, or only to echidnas specifically? As far as I know, Fodor gives no further details about this.

doesn't mean that the concept was there just waiting to be named. In learning what "atom" means, I'm probably *simultaneously* forming the concept ATOM. My niece, in listening to what I'm telling her about echidnas, is somehow forming a new concept, likely on the basis of concepts that she already had. The learning of some predicates may trigger the formation of a new concept (by processes to be explained by psychology), but there is no need to assume that concepts are always prior to words in a temporal sense. Learning what a word means may require forming a new concept, so learning a predicate would only happen when a concept is attached to it, but it doesn't require having that concept at any moment prior to the learning of the predicate. In the case of ECHIDNA, it only appeared because of the linguistic interaction.

Also, this kind of case, where concepts are acquired as a consequence of linguistic communication, speaks not only against Fodor's view on predicate learning, but also against his view of concept acquisition. As we've seen, Fodor argues that there isn't really such a thing as concept learning. That is because he conceives concept learning as an inductive process, which involves formulating hypotheses about the extension of a concept. If one is formulating hypotheses about what things belong to the extension of a concept, one already needs to have that concept. The conclusion is that concept learning, so conceived, is an inherently circular process. But whatever cognitive processes are involved in my niece's learning of the concept ECHIDNA, it seems unlikely they have the form of formulation and confirmation of hypotheses.

As we've seen, in *LOT 2* Fodor no longer adopts a strong nativism about primitive concepts. Rather, he suggests that concepts can be acquired (but not learned) and that concept acquisition may proceed in two stages: first we learn a prototype, and then, by neurological, non-cognitive processes, we acquire the concept whose prototype we learned. But, this account doesn't seem right either as an explanation of my acquisition of the concept ATOM, nor as an explanation of my niece's acquisition of the concept ECHIDNA. The first problem is that these concepts don't seem to have prototypes. But even if they have, their prototypes could conceivably have been learned after the acquisition of the concept. My niece can acquire the concept ECHIDNA before, and maybe without ever, learning its prototype. The second problem is that, even if there is a neurological or biological explanation for what happens in the acquisition of those concepts, which there certainly is, that explanation can hardly be sufficient to account for all that happens in the acquisition process. My niece learned about echidnas because I told her about echidnas, and not, or not primarily, because her neurons behaved in such and such a way.

I learned about atoms because my physics teacher taught me about them. To the extent that linguistic communication involves psychological processes, at least part of the process of acquiring those concepts will involve reference to psychological processes. As Margolis and Laurence note, in a paper challenging Fodor's views about concept learning and concept acquisition, "many concepts are learned via the operation of psychological processes that go beyond stereotype formation." (2011, p. 536). Most of the concepts we have, both primitive and complex, were acquired by cognitive processes that are more aptly described by psychology and not primarily by neurology, and, as Margolis and Laurence point out, these processes can legitimately be characterized as concept learning, even if not always in the sense of hypothesis formation and confirmation that Fodor favors.[162]

In any case, Fodor could perhaps modify his view about natural language learning, conceding that learning what a predicate means eventuates in the acquisition of a concept which is coextensive with the predicate, a concept that the person didn't have prior to the linguistic interaction. This would demand a revision both of the idea that concept acquisition is a noncognitive process that proceeds via stereotype formation, and of the idea that learning a predicate requires having the concept which the predicate is an expression of. He could then adopt the idea that language plays a mediation role between properties in the world and symbols in the language of thought. That is still compatible with the denial of a cognitive conception of language, according to which some thoughts and concepts have natural language expressions as their vehicles. Sure, without language I wouldn't come to think about atoms, and language definitely played a causal role in my niece's *acquisition* of the concept ECHIDNA. But that doesn't mean that language determines the *identity* of the concept, or that it is thereafter constitutively involved in her thoughts about echidnas. Words are perhaps part of the causal chain that goes from properties in the world to my acquisition of certain concepts, but they are not essential for the reference relation to hold. I think, though, that for some concepts, such as color concepts, a somewhat stronger role needs to be attributed to language – not simply the role of being a link in the causal chain that goes from properties in the world to concepts. That's what I'm going to argue for in the next sections.

---

[162] Also, I think it is reasonable to assume that ECHIDNA and ATOM are primitive concepts, that is, concepts that are not constituted of other concepts. Showing that they can be learned, or acquired by psychological processes, would challenge the idea that "where a theory of concepts draws the distinction between what's primitive and what's not is also where it draws the distinction between what's innate and what's not." (FODOR, 1998a, pp. 27-8).

8. The case of colors

We have seen that Fodor, despite some occasional remarks, thinks that the language we speak cannot influence in any strong sense the primitive concepts we have. He holds that concepts that are expressed by monomorphemic words are usually primitive, that is, internally unstructured. Initially he thought that those concepts had to be innate, whereas he now seems to concede that at least some are acquired. But still, language doesn't play any role in determining the identity of a concept. Concepts are partly individuated by their extensions, and reference is a semantic property that belongs primarily to concepts. To the extent that words refer to things, it is only via the reference of thoughts. What a concept refers to is determined by causal connections to the world. Even if Fodor accepts that the acquisition of some concepts is mediated by language, what a concept refers to has nothing to do with language. A concept means what it does, or has the extension that it does, because of causal relations between the mind and properties in the world (regardless of whether language is part of the causal relations). As for the semantics of predicates of natural languages, it is supposed to be completely derived from the semantics of thought.

Even if Fodor no longer says that most of our concepts are innate, I suspect that he would still hold this view for color concepts. That is because colors are properties that people with normal color vision naturally experience, unlike atoms or echidnas.[163] So they are the perfect candidates for being innate concepts – again, innate in the sense of being simply triggered by experience, as a consequence of innate perceptual mechanisms. Presumably, the more mediation you need to acquire a concept, the less likely that concept is to be innate. And for colors, supposedly all that is required is the opening of our eyes.

What I intend to explore in this section is the idea that Fodor might have to give up some of these views in light of one empirical observation: that there can be, as it is well known by now, variations in some categories admitted by different languages. The example I'm going to use here concerns the differences in how languages divide the color spectrum, because color concepts are the best candidates for innate concepts, and because the mutual influence of color words and

[163] I won't be concerned here with the issue of whether colors are really properties of objects or whether they are mental properties. But I think the case of color is analogous to the case of heat. Just as physics gives a physical description of heat, which makes no reference to the sensation of heat, so can color be described in a purely physical way, as the reflectance of light from a surface. In this sense, colors are really there in the world. The qualitative experience we have results from our eyes and brain's detection of the reflectance of light in the objects.

concepts is relatively well studied. But I suspect that the analysis I will offer here could be extended to other concepts as well.

Just to be clear, I think Fodor is right in saying that the language we speak does not determine what we *can* think about (and therefore that a radical linguistic relativism is probably false). It seems that we can always make certain distinctions and create new categories that are not directly expressed by words in our native language. We have an incredible capacity of making concepts out of anything, which is manifested by the constant addition of new words in the dictionaries. So not only do we constantly combine the concepts we have in new ways (as we've seen in the productivity argument), we also seem to have a capacity to keep adding concepts to our repertoire. Since we can think about things for which we have no words, presumably we have more primitive concepts than words.[164] As we have seen in section 3, some phenomena would be left unexplained if we were to assume that thought entirely depends on language. We are not, so to speak, locked inside our languages. But a weaker claim, according to which the language we speak may influence the way we in fact think (and some of the concepts we have), or the idea that the semantics of a predicate may influence the semantics of a concept, should be investigated and cannot be easily dismissed.

To be sure, there is a sense in which it seems right that words are merely there to convey thoughts, and that they mean whatever they do because of what concepts they express. For instance, it seems that we would be capable of thinking about certain things even if we had no words to refer to them. Everybody accepts that there is a high degree of arbitrariness in the sound or gesture that is used to represent a certain object or property. Portuguese speakers use the word "maçã" to refer to apples, English speakers use "apple", French speakers use the word "pomme", users of Libras, the Brazilian sign language, make a gesture similar to the movement we make while holding and eating an apple, and so on. The impression we have is that there is a kind of object in the world, namely apples, about which we could think independently of the use of words, and to which we can refer in different, arbitrary ways that depend on the tacit conventions among the speakers of a language. Unlike the word we use to refer to apples, our APPLE concept, seems to be shared by speakers of different languages – and presumably also by individuals who do not speak any language, and by members of other species, which eat apples. A child can

[164] See, for instance, Li et al. (2011). What they show is that speakers of Tzeltal have no problem in solving tasks which presuppose the use of egocentric spatial concepts, even though their language doesn't lexicalize the distinction between left and right (which are part of the egocentric frame of reference).

certainly desire, or demand to eat an apple even before she has mastered the word "apple". So some predicates of natural languages seem to do nothing more than express, in initially arbitrary, but afterwards regular ways, ideas that exist independently of a particular language. The word "maçã" means *apple* because Portuguese speakers use this word with the intention of talking about apples. And that word would not exist were we not able to think about apples. This suggests that concepts are prior to words, and that the predicates of a language mean what they do because they are used to express certain concepts. "Maçã", "apple", etc., are mere sequences of sounds that in some sense only come to life because they accompany thoughts about apples. So in this case it does seem reasonable to assume that the semantics of "apple" is derived from the semantics of APPLE.

Apples are also a good example of the triggering of concepts by experience. It seems reasonable to assume that being exposed to apples can trigger in me the attainment of the concept APPLE. Maybe Fodor is even right that APPLE means *apple* because apples cause APPLE, and that any token of APPLE not caused by apples is asymmetrically dependent upon those that are. If a tomato caused a token of APPLE, APPLE would still mean *apple*, because tomatoes would not have caused APPLE tokens had apples not caused APPLE tokens.

Fodor's view on the priority of concepts over words, as well his causal-referential view about the semantics of concepts seem to work well for concepts like APPLE. But the mind does not live on APPLES alone. I've indicated earlier that some of our concepts, those which are acquired by means of language, challenge at least Fodor's views on predicate learning, and the idea that concepts are prior to words in a temporal sense. I think they also challenge Fodor's extreme nativism, but Fodor could object that he no longer holds this view. Still, I think there are other examples of concepts that are even more worrisome for Fodor's views, namely, lexical concepts (expressed by monomorphemic words) that are not coextensive across different languages. I'm going to focus here on color concepts.

There is clearly variation in color terminology across different languages.[165] Some languages have as few as two basic color terms (that mark the opposition between light and dark

[165] Other differences among languages that have been studied concern variations regarding the names of body parts, names for kitchen utensils, kinship terms, and names for different types of movement. Verbs like "hop", "skip" and "stroll", for instance, don't have direct translations in Spanish or Portuguese, for example. I think it is no coincidence that we find the most variation across languages in names for things that are not clearly demarcated in nature.

colors).[166] There are several languages that, for example, do not mark the distinction between the colors blue and green, but instead use the same word for both colors. That is the case of Tzeltal, a Mayan language spoken in Mexico, of Himba, an indigenous language spoken in Namibia, and of Berinmo, a language spoken in Papua New Guinea. These last two have only five basic color terms. Also, some languages, like Russian and Greek, have different monomorphemic words for light blue and dark blue. To be sure, color perception is innate. How many distinct colors we can perceive is determined by our biology: people in different places and cultures are able to perceive the same differences among different colors, regardless of how many basic color terms their language has. That is, humans with normal vision, regardless of their culture or language, are capable of making the same judgments about whether two hues are identical or different. According to Roberson and Hanley (2010), we can discriminate approximately 2 million colors. But still, color terminology varies a lot across languages, and as I will try to show, this fact, along with some empirical findings, doesn't seem compatible with at least some of Fodor's views.

It is worth noting that, against an extreme linguistic relativism, Berlin and Kay (1969) have argued, in a classic study, that the lexical coding of color is not as arbitrary as some linguists thought at the time. According to Berlin and Kay, color naming follows a certain pattern of evolution (with names for black, white and red appearing before names for blue, for instance). They also claimed that, when asked to indicate which color is the focus, or best example, of each of the color category they have a word for, speakers of different languages tend to choose colors in predictable areas of the spectrum. When presented with 329 color chips with different hues and brightness, speakers of different languages chose chips that were relatively close together as the most representative for a given color term in their language. This suggests that different languages don't cut the spectrum in a completely arbitrary and untranslatable way, as a strong linguistic relativist would suggest, but instead, according to them, there are constraints, possibly dictated by biology, regulating the variation of color terms across languages.

Still, the number of basic color terms that a language has varies, and consequently, the fewer color words a language has, the more inclusive those words will be. The boundaries for each color category are not the same across different languages. So color words have different

[166] The notion of basic color term is introduced by Berlin and Kay (cf. 1969, p. 5-6). In order for a term to count as a basic color term, according to Berlin and Kay's criteria, it needs to be monolexemic (excludes "lemon-colored"), to refer to a color that is not a subset of a broader category (which excludes "crimson"), not be restricted to a particular class of objects (such as "blond") and refer to something that is psychologically salient and can be shared by other speakers (which excludes "the color of the rust on my aunt's old Chevrolet").

extensions in different languages. They cut the spectrum according to different boundaries. Even if Berlin and Kay are right in saying that the colors taken to be focal are similar across different languages, what they take to be focal are not individual colors, but rather still relatively large regions in the spectrum. For instance, out of forty Tzeltal speakers they consulted, thirty-one located the focus, or best example, of "yaš" (their term for blue-green) in the green area of the spectrum, whereas nine located it in the blue area (BERLIN; KAY, 1969, p. 11). Even if there are constraints regulating color naming, there is still variation both in the extension of color categories, and in what colors individuals take to be the best example of each of their linguistic categories.

How then, according to Fodor, should we deal with the case of a language that, like Tzeltal or Himba, doesn't lexicalize the distinction between blue and green? Do a speaker of Himba and I have the same color concepts? Do we conceive colors in the same way? When I think that my favorite color is green, and when a speaker of Himba thinks that his favorite color is *burou* (the word they use for both blue and green), do we have the same thought? If I think that the sky is blue, and a speaker of that language thinks that the sky is *burou*, are we having the same thought? When we think about how many colors there are, do we reach the same conclusion?

If we accept, as Fodor suggests, that concepts are individuated in part by their extensions, and that words express primitive concepts, it seems that we should say that a speaker of English does not always have exactly the same thoughts as a speaker of Himba does about the colors blue and green, for they have primitive concepts with different extensions. But, as we've seen, primitive concepts are supposed to be innate – especially color concepts. And how could different people have different innate color concepts? If these color words are really expressions of different primitive concepts, then it seems that these concepts cannot all be innate at the same time. Also, if Fodor accepted that speakers of different languages have different color concepts, it seems that he would also have to accept that the language we speak does have some influence on the primitive concepts and the thoughts we have, and maybe that the semantics of thought can be influenced by the semantics of language.

The fact that color terminology is not coextensive across languages creates a problem for the consistency of 3 of Fodor's ideas: (1) words express primitive concepts; (2) primitive (in particular color) concepts are innate; (3) the semantics of words has no influence on the semantics of concepts. Fodor has two main alternatives. The first is to abandon the view that color words

express primitive concepts, in order to preserve his views on the innateness of primitive concepts, and the independence of the semantics of thought from the semantics of language. The second is to stick to the idea that color words express primitive concepts, at the price of abandoning nativism about color concepts and conceding that the language we speak may influence the semantics of some primitive concepts we have.

I think only the second alternative is viable, but let's explore the first alternative. A possible way out of a weak linguistic relativism would be for Fodor to say that color words, unlike many others, sometimes express complex concepts (in the case of "burou", two syntactic units each with a different extension), and not one primitive concept (one syntactic unit whose extension is, from our perspective, disjunctive). Maybe what gets tokened in the heads of Himba speakers when they think and talk about *burou* is a complex concept, BLUE OR GREEN. This would mean that they have the same primitive color concepts that we have, but form a complex color concept that we typically don't form.

Likewise, he could say that their word "burou" is sometimes ambiguous. Fodor could say that just as Himba speakers can token these two concepts disjunctively, sometimes they will have in mind either BLUE or GREEN, but will express these concepts ambiguously in their language. So they in fact have two concepts, one whose extension is the same as the extension of the word "blue" and another concept whose extension is the same as the extension of the word "green," but this distinction between concepts is not captured by their language. Even if they use only one word to refer to both blue and green, there could be a distinction that is drawn on the mental or conceptual level. So even though a Himba speaker might say that his favorite color is *burou*, which is ambiguous between green and blue, his thought will not be ambiguous. He will be thinking either BLUE or GREEN (or even a disjunction of these two concepts) when he thinks about his favorite color, or when he thinks about the color he wants to paint his house. What happens to the speakers of Himba when they talk about colors would be something similar to what happens to me when I talk about banks. If I say "I'm going to the bank", my sentence is ambiguous because I can use it to communicate either that I'm going to the river bank or that I'm going to the financial institution. But my thought is not ambiguous: I presumably have two concepts for the two possible banks, and I use one or the other when I think that I am going to the bank. It is just accidental that the words that express these concepts in English are homophones.

This would in fact be a natural way out for Fodor, because he thinks that natural languages, unlike thoughts, are quite often ambiguous. He favors a view according to which thought is more complete than natural language (cf. 2001a). So he could say that, normally, words express primitive concepts – so there is a more or less straightforward correspondence between words and concepts. But that correspondence is not perfect. Sometimes the same word (individuated phonetically) can be used to express more than one concept, because it is ambiguous, and sometimes a word will be used to express a disjunctive concept. Both could be the case for color concepts.[167]

Against this, though, it could be said that whereas the financial institution and the river area are clearly distinct entities in the world, the same is not true for blue and green. They form a continuum on the color spectrum, so there is at least some plausibility in the assumption that Himba speakers simply cut the spectrum in a different, more inclusive way than English speakers – instead of assuming that we all share the same primitive color concepts, only expressing them differently. More importantly, there is in fact some empirical evidence in support of the view that color words are not ambiguously expressing universal concepts, but are, instead, a good mirror of our broad color concepts. Some studies show that differences in linguistic categories correlate with differences in discrimination, memory and perception. Let's briefly consider some of these.

Roberson et al. (2000, 2005) performed a series of experiments with Himba, Berinmo and English speakers. Both Berinmo and Himba have only five basic color terms (English has eleven) and neither language marks the distinction between blue and green. Their color categories are similar, though not exactly coextensive. Some Himba categories have different boundaries as compared to Berinmo categories, and the colors that are judged to be focal, or best examples of each category, vary across languages (even if obeying certain general constraints, as Berlin and Kay claimed). In one experiment, Roberson et al. (2000, p. 392, 2005, p. 399) showed speakers of Himba, Berinmo and English one target color chip, removed it from sight and, after 5 seconds, showed participants two color chips, one that was the same as the target, and one that was different. Participants were then asked to point to the same chip they had seen before. What they

[167] The claim that color categories are universal, despite differences in color terminology, would not be unique to Fodor. Heider (1972) made some experiments with Dani speakers, a language that has only two basic color terms, and found that colors that were focal for English were better recalled by Dani speakers, even if they lacked a term for them. This suggests that we group colors into categories around the same hues, even if we lack terms for them. However, Roberson et al. (2000, 2005) failed to replicate the findings with Berinmo and Himba speakers, for they found significant differences between them and English speakers, as we will see in what follows.

found was that participants were more accurate when the colors of the two chips belonged to different categories in their language than when the colors belonged to the same category. And whereas English speakers were more accurate at recognizing color chips when the alternatives crossed the blue/green region, that effect was not observed in Himba or Berinmo speakers. Berinmo and Himba speakers performed better when the two color chips belonged to different categories in their own language, and blue and green are not different categories for them. If they categorized blue and green as different colors, despite their language lacking that distinction, we would expect Himba and Berinmo speakers to have a similar performance in the experiment to English speakers, but they didn't.

In another experiment Roberson et al. challenged Berlin & Kay's (and Heider's) assumption that focal colors are universal. According to them, not all colors that are considered focal for English speakers are considered focal for Himba and Berinmo speakers, and vice versa. Some hues that English speakers classify as pertaining to the boundary of a given color category (not in the center of the category), or as being internominal (falling between two categories) are actually focal for Himba speakers, and some that are focal for English speakers are not focal for Himba speakers. In one experiment (2005, p. 389), participants were shown one color chip for 5 seconds and, after a 30 second interval, they had to indicate which color, of a 160-chip array, they had just seen. Some of the target color chips they were shown were of focal colors for English speakers, and some were of boundary or internominal colors for English speakers. Whereas English speakers had a better performance when presented with focal colors in their own language than when presented with colors that were internominal or at the boundary of a category, Berinmo and Himba speakers didn't perform better when the color chips were focal for English. Rather, they showed a better recognition for other colors. The natural interpretation, then, is that their focal colors are different, and thus that there are no universal focal colors. Speakers of a language with a grue category don't have two focal points, one for green and one for blue. Their focal colors are compatible with their color vocabulary.

As I mentioned earlier, the Russian language has one word for lighter blues ("goluboy") and a different one for darker blues ("siniy"). That is, Russian, unlike English, forces their speakers to specify whether one is talking about dark blue, or light blue. In an experiment by Winawer et al. (2007) English and Russian speakers were instructed to indicate as quickly and accurately as they could which of two bottom squares had the same color as a square that was

presented on top. They found that "Russian speakers were faster to discriminate two colors if they fell into different linguistic categories in Russian (one siniy and the other goluboy) than if the two colors were from the same category (both siniy or both goluboy)" (WINAWER et al., 2007, p. 7783). The same category advantage, or categorical perception effect, was not observed in English speakers (even though, when English speakers were asked to divide the blue stimuli into the categories "light blue" and "dark blue", their categories closely corresponded to the categories of "goluboy" and "siniy", respectively). So the authors concluded that linguistic categories can influence perceptual performance.[168] What is ultimately being suggested is that colors that belong to the same linguistic category are perceived as being more similar than colors that belong to two distinct categories, even when they are equally spaced according to a given color metric system, such as Munsell's. If this finding can be generalized, we would expect to observe the same category advantage in English speakers regarding their perceptual discrimination of blue and green, over speakers of Himba and Berinmo regarding their discrimination of the same colors.[169]

These differences in memory and discriminatory capacities suggest that speakers of different languages don't conceive colors in the same way after all – they don't carve the color spectrum in the same way, around the same focal colors. The difference in performance between Himba and English speakers suggests that Himba speakers don't have two distinct concepts, BLUE and GREEN, which are expressed either as a complex, disjunctive concept, by the word "burou", or ambiguously, but rather that they have one syntactic concept with an extension that is disjunctive from our perspective. Likewise for the differences in concepts between Russian and English speakers.

To be clear, I'm not rejecting the idea that sometimes color words are ambiguous, or polysemous. I may have a particular hue in mind when I say "blue", a narrower region of the spectrum and not the whole part of the spectrum that I would also call "blue". We can create color concepts on the fly. I can for instance think about the blue of this car, and plan to paint my house the same color. In these cases, I presumably have more specific color concepts in mind,

---

[168] It is interesting to note, though, that the experiment by Winawer et al. did not show that Russians were faster to discriminate between lighter and darker blues than English speakers. On the contrary, English speakers were overall faster than Russian speakers in all the discriminatory tasks, despite lacking specific linguistic categories for each type of blue. The authors speculate that this difference might be due to Russian speakers being less experienced in using computers or in taking part in experiments than English speakers.

[169] For another experiment on related topics, see Kay and Kempton (1984). They found that "the English speaker judges chip B to be more similar to A than to C because the blue-green boundary passes between B and C, even though B is perceptually closer to C than to A" (Kay and Kempton, 1984, p. 77). The same effect was not observed in speakers of Tarahumara, which is another language that has a grue term.

concepts which don't include the whole extension of BLUE, but whose extension is a subset of the extension of BLUE. These more specific color concepts may be shared by speakers of different languages. But still, I do sometimes conceive blue as encompassing a certain region of the spectrum that is located between green and purple. If a friend tells me that she painted her house blue, I might expect anything from light to dark blue. What categorical perception effects show is that the grouping of the spectrum in these broad categories has some impact on the way we perceive and recall colors. A natural suggestion, then, is that these effects are explained by the assumption that speakers of different languages have different general color concepts.

As far as I know, Fodor does not address this type of example. He says, though, that our discriminatory capacities do not interfere with the contents of our concepts. He illustrates this point by saying that even though he is not capable of distinguishing between elms and beeches, it does not follow from this that he has only one concept whose extension includes both types of trees. According to him, "when I think *elm* I am in a computational state that is, to all intents and purposes, functionally indistinguishable from the one I'm in when I think *beech*" (FODOR, 1994, p. 33). So, from a computational point of view, both concepts are syntactically identical for him, since they have the same causal powers – they produce identical behaviors and have the same role in thought processes. Identity of causal powers is sufficient for identity of syntax, but not necessary. But since Fodor thinks that concepts are also individuated by their extensions, or their referents, from a semantic point of view, ELM and BEECH are different concepts – they refer to different properties in the world: "broad content individuation insists, however, on distinguishing between these states since they have, by assumption, different truth conditions" (FODOR, 1994, pp. 34).[170] Roughly, Fodor thinks that semantics is not to be confused with epistemology, nor, surprisingly, with psychology. From the point of view of the semantics of concepts, it is irrelevant that I myself do not make certain distinctions. My actual abilities do not interfere with the contents of the concepts I have. According to Fodor,

> Semantics, according to the informational view, is mostly about counterfactuals; what counts for the identity of my concepts is not what I *do* distinguish but what I *could* distinguish if I cared to (inter alia, what I could distinguish by exploiting instruments and experts). (…) According to externalism, having the concept ELM is having (or being disposed to have) thoughts that are causally (or

[170] This view is problematic. It has, for instance, the undesirable consequence that I can have no idea what some of my thoughts are about. We shouldn't have to multiply concepts whenever sciences multiply properties. But I won't discuss this issue here.

> nomologically) connected, in a certain way, to instantiated elmhood. Punkt. It is, to put the point starkly, the heart of externalism that *semantics isn't part of psychology*. The content of your thoughts (/ utterances), unlike, for example, the syntax of your thoughts (/utterances), does not supervene on your mental processes. (FODOR, 1994, p. 37).

Returning to our example, Fodor might then try to insist that even though Himba speakers use "burou" to talk about green and blue, and that they may be less efficient in distinguishing blue and green than they are in distinguishing any other two colors for which they have two different names, it does not follow that they have a primitive concept whose extension is both blue and green.[171] If we take blue and green to be different properties, and we take concepts to be individuated in part by their extensions, Fodor maybe would have to say that Himba speakers have in fact two distinct primitive concepts, BLUE and GREEN, since BLUE and GREEN have different extensions, or express different properties. This way, it could be said that we all have the same color concepts after all.

The problem is, if Fodor said that there cannot be a unique BUROU concept because blue and green are different properties in the world, and that the word "burou" expresses in fact two primitive concepts, BLUE and GREEN, we might ask what is to count as a color property. Since we are able to discriminate an extremely large number of hues, why should we take BLUE and GREEN to be primitive concepts, and blueness and greenness to be basic color properties, so to speak? There seems to be no reason not to consider light blueness and dark blueness also as different properties (as Russians do), in which case "blue" should be presumably ambiguous. But then there seems to be no good reason to stop there either. Why not distinguish pale pink from bright pink, light green from dark green, etc.? If Fodor wanted to preserve the objectivity of color properties, in order to preserve the objective aspect of the semantics of thought and free it from the influence of the particular language we speak, it seems that he would be forced to hold an extreme view, according to which we have a primitive concept for each of the hues that we are capable of distinguishing (unless some independent standard is suggested for determining which

---

[171] There is a relevant difference between the two cases, which is that Fodor is in fact unable to distinguish between elms and beeches, whereas Himba speakers are perfectly capable of distinguishing between blue and green, just like English speakers are able to distinguish between light and dark blues. It is just that they are presumably slower in doing so than they are in distinguishing between colors for which their language have different words. But the point here is that Fodor takes extensions to determine the contents of concepts. So if there are two distinct properties in the world, blue and green, it seems that he would have to say that Himba speakers have two different concepts, regardless of what their discriminatory capacities may be like.

color concepts are primitive). But that doesn't seem realistic, from a psychological point of view. We would have to postulate thousands – or even millions – of primitive color concepts, whereas the empirical evidence suggests that we group different hues into broad categories. If we accepted the assumption that there is a concept for every hue of color we can distinguish, the categorical effects on perception and memory would be left unexplained.

I have been exploring a way in which Fodor could avoid being committed to the idea that Himba speakers and English speakers have different color concepts, which could possibly force him to accept a form of linguistic relativism. I said Fodor could try to abandon the idea that color words in different languages express different concepts, at least for some color words. He could say that there will be a different concept for each color property in the world, regardless of one's language or discriminatory capacities. Speakers of Himba would have the same concepts we have, the concept BLUE that is triggered by blueness, and the concept GREEN that is triggered by greenness. But, as I argued, that doesn't seem to be a very promising strategy, in part because it is not clear what the color properties are. The issue here seems to be that color properties don't have the same conditions of individuation as natural kinds and concrete objects in the world. In the case of elms and beeches, botany tells us that there are two distinct natural kinds in the world, two discrete different kinds of entities, about each different generalizations can be made. But it is not obvious how color properties should be individuated. The same kind of deference to experts, typical of natural kind properties, doesn't usually apply in the case of colors. Colors, unlike trees, are not discrete entities in the world, but form a spectrum and are, to some extent, mind dependent.[172] What seems to be objective and approximately universal is the number of hues we can distinguish, but to assume that we have a concept for each hue we can potentially distinguish seems to be an unnecessary multiplication of entities – not to mention that, as I said, it leaves categorical perception effects unexplained.

Assuming we should not appeal to different color properties in order to individuate the contents of color concepts, the natural alternative, more plausible from a psychological point of view, is to hold that color concepts have the same extensions as the words that express them. I think that color concepts make actually a good case for the idea that lexical items express

[172] In *Concepts* (p. 135) Fodor says that colors are appearance properties.

primitive concepts.[173] The best way to know what someone's primitive color concepts are is to know what language that person speaks.[174] BUROU is a primitive concept for Himba speakers, whose extension for English speakers corresponds, approximately, to the extension of BLUE plus the extension of GREEN. Assuming extensions are part of what individuates concepts, we have different color concepts from Himba speakers, because our primitive color concepts have different extensions.[175]

But now we are back to where we started. If Fodor were to preserve the idea that color terms express coextensive primitive concepts, he would have to say that speakers of English, Himba, and Russian (and, for that matter, speakers of all, possibly thousands, of languages, whose color terms are not mutually coextensive) have different primitive color concepts (for they have different basic color terms). So in this scenario, Fodor would have to acknowledge that, strictly speaking, speakers of English and of Himba do not always share exactly the same thoughts about colors, for they have concepts with different extensions (and extensions individuate concepts). He could still say that their thoughts would be sufficiently similar to preserve most of the psychological generalizations with respect to colors. But still, if Fodor allowed that there is this difference in primitive concepts, it seems that he would have to abandon the view that color concepts are innate. Also, he would have to accept that the language we speak influences what we think about, or that the semantics of the language we speak may be (in some sense) prior to the semantics of thought. Now it seems that Fodor cannot hold both that words express primitive concepts and that color concepts are innate, or that the semantics of thought is prior to the semantics of language, or even that natural languages have no semantics. If speakers of different languages have different primitive color concepts, what else could explain this difference if not that the language one speaks can, to some degree, determine the content of those concepts, and shape their boundaries?

---

[173] Here I'm using "primitive" in the sense of not being decomposable into simpler concepts that define it, and not in the sense of being unstructured. I leave it open that color concepts could be structured around the prototypical color of each color category.

[174] This is not to say that the program on lexical decomposition is destined to failure. It may well be that some verbs and nouns are in fact expressions of complex mental representations. My point is simply that this is not a reasonable assumption for color concepts.

[175] Again, this is not to say that narrower color concepts cannot be formed. But I'm assuming that English speakers' concept BLUE is probably not constituted by any other more primitive concepts (e.g. GOLUBOY and SINIY). It has the same extension as the word "blue", and is not itself a disjunction of more primitive concepts. I concede, though, that color concepts may be structured around prototypes, or the best examples of each color category.

One strategy Fodor could adopt to try to avoid this conclusion is to minimize the influence that language itself has on thought. Fodor might say that perhaps Himba speakers initially formed a comprehensive concept that includes in its extension both blue and green because it was not useful for their community to draw this distinction. Perhaps they were not surrounded by that many blue things, so there would be no use to create a different word to refer to blue. They then used the word "burou" to express a concept that has in its extension both green and blue, which was good enough for their needs. Teaching this word to future generations perhaps reinforces that there is no need for more precise concepts. In addition, Fodor could insist that, for the word to be learned, it would still be necessary for a child to already have a concept with that extension. Fodor could say that the word only names a concept that is already present, perhaps originating from what the experience of people in their community has shown to be relevant. Perhaps the way the words of our language are used to categorize objects can force us to make certain distinctions more often than speakers of other languages. But it is possible that what influences our thinking is not the words pure and simple, but the experience and the training we are given to making certain distinctions. Maybe Russian speakers are faster at discriminating cross-category than within-category blues not merely because they have two completely different words for light blue and dark blue, but rather because they were trained early on to make those distinctions, with more emphasis than English or Portuguese speakers. Experts in trees know how to distinguish elms from beeches primarily because they paid attention to their differences, or were trained to see their differences, not necessarily because they have different words for elms and beeches. Perhaps Russians are "experts", so to speak, in blue, just as some people are experts in distinguishing types of trees. Words can help fix distinctions, and can be correlated with the time one takes to perceive certain differences in the world, but Fodor could say that words are not the primary cause of our concepts having the extensions that they have, or of us being able to make certain distinctions faster than others.

Pinker adopts precisely this line of reasoning, trying to minimize the influence that the language we speak has on our concepts. He discusses the popular idea that Eskimos have a larger number of words for snow than we find in English or Portuguese (cf. PINKER, 2007). What a proponent of linguistic relativism might say is that having more words for snow leads them to pay closer attention to snow than we do. Pinker notes that this explanation reverses cause and effect. It is more reasonable to suppose that Eskimos have more words for snow (if they do, which is still

debatable) because they live in an environment where paying close attention to snow is important and useful, whereas speakers of a language like Portuguese presumably have fewer monomorphemic words for snow, because snow is not part of their environment, so it is not particularly relevant to make those distinctions. That is, it is not the language that causes Eskimos to notice different types of snow, but their environment and interests. Their different words for snow are a consequence of the concepts that originated from the close attention they pay to snow, not the other way around.

There is no need to deny that words only appear in the first place because of human interests and needs. But it is also true that, in some sense, in saying that Russian children have the concepts SINIY and GOLUBOY because they are trained to completely differentiate light blue from dark blue we would only be pushing language's influence on thought one step back, rather than avoiding it completely. That's because Russian children are being trained to completely distinguish dark blue from light blue, not taking them to be two examples of the same broad color category, mainly because their parents are teaching them Russian, and the Russian language has two different, monolexemic words, for each part of the spectrum. An American child is not taught to distinguish light blue from dark blue with the same emphasis as it is taught to distinguish between green and blue, because that distinction doesn't appear in English's basic color terms. Here the case of color differs from the case of snow in that different cultures can be surrounded by the same colors and still end up carving up the color spectrum in different ways. Assuming Russians don't have radically different color experiences than Americans, and that the culture is relatively similar, it seems reasonable to assume, I think, that language, and not only non-linguistic cultural factors, is playing a big part in the formation of color concepts in children. It is influencing which color concepts one is going to end up having, and indicating how the color spectrum is supposed to be divided. Unlike the case of words for snow, words for different types of blue are not obviously something that an environment strongly suggests. There are some distinctions, though, that can be introduced to help us in considering the ways in which we can conceive the issue of the priority between language and thought. They will be introduced in the next section.

9. Priority of the semantics of language over the semantics of thought

As we've seen, Fodor often says that the semantics of thought is prior to the semantics of language. If it is true that speakers of different languages have different color concepts – if language is causing our color concepts to have the boundaries that they have – then it seems that the semantics of language can have some influence on the semantics of thought. But it is important to clarify that there are different ways in which we can understand Fodor's claim that the semantics of thought is *prior* to the semantics of language. Partly following Devitt (2006), the priority can be ontological, explanatory, temporal or causal. It is not always easy to disentangle all these kinds of priority. It is also useful here to add the distinction between two perspectives that we can take on this issue: a *historic* and a *developmental*.

First of all, I think it seems right to assume that sentences and words wouldn't mean anything if it weren't for thoughts, but thoughts would still represent or mean something even in the absence of language. Nothing I have said so far challenges that. So in this sense, which can be called ontological, the semantics of thought is prior to the semantics of language.[176] Sentences are about things in virtue of conventions, whereas thoughts naturally have the property of being about things. In addition, in a related sense, we don't need to mention language when saying what the content of a concept *is*. As Fodor puts it, "the Mentalese story is that the content of thought is *ontologically* prior to natural-language meaning. That is, you can tell the whole truth about what the content of a thought is *without saying anything whatever about natural-language meaning*, including whether there is any." (1998b, p. 68). So language is not part of what constitutes the content of a thought or concept.

Saying that thought is ontologically prior to language has temporal implications, as Devitt (2006, p. 130) notes.[177] A child that is incapable of thinking won't learn a natural language – from a temporal point of view, it needs to be capable of thinking before it can learn to talk. And, as Devitt and Carruthers note, language wouldn't have appeared in our species if hominids were incapable of thinking. But here I'm not concerned with thought in general. I am considering how

[176] Here I'm partially following Devitt's terminology. He argues that linguistic competence is constituted by a conceptual competence and a processing competence. For him, "the conceptual competence has an 'ontological priority' over the linguistic competence: it is metaphysically possible to have the conceptual competence without the linguistic competence, but not vice versa" (2006, p. 130).

[177] And Fodor himself, when he says that "we can reasonably expect, having once answered the question about the relative priority of thought and language in the explanation of content, that connected issues about which comes first in the order of ontogenetic or phylogenetic explanation should then fall into place." (2001a, p. 02)

concepts relate to words, in particular color words. So I'm not asking whether a child needs to be able to think before it can learn how to talk – the answer being clearly yes. I'm asking whether a child's concepts are coextensive with the predicates of the language it is learning, before it learns it.

If we consider words from a historical perspective, it is reasonable to assume that they are usually introduced to express concepts. Color words were introduced to express color concepts – so concepts came first in this sense. So I agree with Pinker (2007) that the appearance of new words is a consequence of particular ways of conceiving the world. It is likely that the different Russian words for types of blue, for instance, initially appeared because of the interests of particular individuals, who had some particular reason to distinguish both colors (hence, thinking of light blue and dark blue as different categories, historically, came before having two distinct words for them). Words are usually introduced to draw distinctions that people, for whatever reason, find relevant.

Fodor also says that the semantics of thought is prior to the semantics of language in order of explanation of semantic content (2001a). What he means by that is that we explain what sentences and words mean by appeal to what thoughts they express, and not the other way around. The idea here is that words mean what they do because they express the concepts that they express. This idea seems to follow naturally from the assumption that the function of language is to express thought, and the idea that the content of thought is ontologically prior to the content of language. Sentences wouldn't mean anything if it weren't for thoughts, and since we use language to express thought, it is only natural that whatever a given sentence means is going to be explained by whatever thoughts speakers of the language use that sentence to express. If we want to explain what a word means, and why it means what it does, we may have to refer back to the thoughts of the speakers that use that word. We will say that the word "blue" means blue because it expresses a certain concept, with a given extension. The case of color concepts does not speak against this kind of explanatory priority of thought over language.

However, if what we want to explain is *why* we have some concepts with the semantic properties that they have, the appeal to the language we speak may be required. If we look at the issue from a *developmental* perspective, "goluboy" and "siniy" are now used to reinforce a distinction that someone thought was relevant. The semantics of some words can be prior to the semantics of concepts at least in a *temporal*, or a *causal* sense, when we consider the issue from a

developmental perspective, that is, from the perspective of a child learning color words and concepts. Just as in the case of the acquisition of concepts through linguistic communication among linguistic competent people, there is no need to assume that children learning color words are simply mapping the color predicates to preexisting concepts. In fact, that is plainly false, as I'm going to argue in the next section. In any case, to the extent that color concepts vary across speakers of different languages, it seems reasonable to assume that language is what *causes* us to have to some color concepts and not others. So words, in a temporal or causal sense, from a developmental perspective, come before concepts – children's color concepts inherit their extensions from the extensions of the words in their language.

I do want to stress, though, that the causal role language plays in the acquisition of color concepts seems to be somewhat stronger than the role it plays in the acquisition of concepts whose referents are clear-cut properties or individuals in the world, like echidnas. In the case of apples, or of echidnas, might serve to connect people to these properties. In the case of color concepts, language also plays a role in causally determining the identity of our color concepts. As Fodor says, concepts are partly individuated by their extensions. If color properties cause color concepts in us, they do so typically via the mediation of language, which serves not only to connect us to independently existing properties, but in this case also to institute that property to us, to establish its limits.[178] Color properties don't come in clear groups in the world; they aren't just waiting to cause color concepts in our minds with a precise extension – like maybe echidnas, or apples. If that was the case, we wouldn't find the variation we find in what people take to be the colors that there are in the world. *Blue*, *green*, *burou*, *goluboy*, are *conventional* properties, for their existence is conditioned to our way of dividing the color spectrum.[179] The different hues that we can perceptually distinguish are somewhat universal and innately specified (though presumably mind-dependent), but the referents of the concepts BLUE, GOLUBOY and BUROU, though perhaps not entirely arbitrary, as Berlin and Kay would argue,[180] are in the end determined by conventions. Given hue X, there seems to be no uncontroversial fact that it will belong to

[178] I don't mean to be saying that color concepts and words have definitive boundaries. Rather, they are great examples of vague concepts and words. But still, they have approximate boundaries, which are, I claim, not given by the mere perception of colors.

[179] Unlike what Fodor thinks, when he says that "there are many properties that are untendentiously mind-dependent though plausibly *not* conventional; *being red* or *being audible* for one kind of example; or *being a convincing argument*, for another kind; or *being an aspirated consonant*, for a third kind; or *being a doorknob*, if I am right about what doorknobs are." (1998a, p. 149).

[180] See also Regier et al. (2010). The authors say there are universal constraints on color naming, but that language can have some influence on the boundaries of the categories.

category Y, and not Z. Where a color property ends, and where it begins, even admitting a certain vagueness to its boundaries, is going to depend on what conventions the speakers of a language adhere to. It seems then that language is somewhat responsible for instituting color properties and for making one's concepts ending up having the extensions that they have. To the extent that concepts are individuated in part by what they refer to, and that what color concepts refer to is determined conventionally, by means of language, language plays a causal role in determining a color concept's identity.[181]

It seems, then, that we have to mention language in order to explain *why* we have color concepts with the semantics that they have, or why we group colors together the way we do. The same sort of explanation is not needed in the case of echidnas, or atoms – even if language mediates the locking to these properties. We can abstract away from language when we explain why we conceive echidnas or atoms the way we do – even if language is a prerequisite for us being capable of thinking about atoms, or echidnas when not in direct causal connection to them. I think about echidnas the way I do not because I learned Portuguese instead of any other language. That's perhaps because echidnas are not conventional properties. But the same is not clearly true of colors. Given that I explain *why* my color concepts have the referents that they have by appeal to the fact that I speak language X and not language Y, there is at least a causal sense in which the semantics of thought is explained in terms of – and is therefore not prior to – the semantics of language (cf. FODOR, 1998, p. 68).

In a sense, we are stuck in a circle here, with thought influencing language and language influencing thought. That's because in order to explain why I have the color concepts I have, I need to mention that I grew up speaking Portuguese, and not Russian or Himba. But in order to explain why we have the word "blue" in English, and why the word "blue" has the semantics that it has, we need to refer back to people's intentions. But the thesis that language plays a causal role in determining a concept's identity is compatible with the thesis that a word's content is explained in terms of a concept's content. So nothing I said speaks against the idea that we explain what the word "blue" means by saying that it means whatever speakers of English use the word to refer to. One can still hold that color words are always used by adults to express their thoughts about colors. So even if our concepts are shaped by these words in the process of learning color

[181] It is possible that language plays this stronger role especially when it comes to classifying qualities, as opposed to substances, or discrete entities.

predicates, there is still an important sense in which the semantics of these words is derived from thought.

My point, then, is simply that language learning plays a role in color concept acquisition, and that we can sometimes legitimately explain why our concepts refer to the properties they refer to by saying that it is because the words in our language refer to what they refer. From a developmental perspective, color concepts can inherit their extensions from the color words of a language.[182] If that's the case, language plays a causal role in the determination of the extension, and therefore of the identity, of color concepts.[183] We use words to express the properties we are thinking about, but we also use words as a means to instruct others how they are to categorize some properties in the world. These are roles that Fodor doesn't attribute to language, and which I think are important to emphasize, even if the circle requires us to continue to explain the semantics of words in terms of the semantics of thoughts.

In stating that the semantics of thought is prior to the semantics of language, Fodor oversimplifies the issues of the relationship between language and thought. Introducing the opposition between a historic and a developmental role of words is useful to make us see that Fodor minimizes the developmental role of language – the role of shaping one another's thoughts. Even though a color term may be initially introduced into a language because of the interests of a group of people – which suggests that, historically, or diachronically, the concept came first, and the word later –, the word can subsequently be used to shape the color concepts of children – to make their concepts similar to the concepts of adults in their community – for they probably wouldn't get to the same ways of broadly dividing the color spectrum only as a consequence of seeing colors in the world. So when we approach the issue from the perspective of a child learning words, the color words she learns are causally involved in the acquisition and determination of

---

[182] If some words have the power of shaping the contents of some of our concepts, it seems that Fodor is not right to suppose, without qualifications, that words merely inherit their contents from thought. So it doesn't seem to be the case that natural languages have no semantics, or that, if they do, their semantics is solely determined by, or derived from, the semantics of thought. The semantics of color terms seems to influence the semantics of color concepts, at least in a causal sense, from a developmental perspective. It would be difficult to explain the causal role that language plays in the shaping of our concepts without the assumption that words themselves have referents, which concepts inherit in the process of language acquisition.

[183] Devitt notes that "the fact that the presence of a word in a language makes it easy for a speaker to gain a concept does not show that, contrary to the Gricean view, the word has a role in determining the nature of the concept thus gained; it does not show that language is prior to thought in the explanation of meaning." (2006, p. 137) I think, though, that for the reasons presented, color words do seem to play a role in determining the nature of the color concepts one acquires.

the identity of their color concepts. The way she conceives colors will be influenced by her observations of adults using color terms.

Finally, even if language is causally responsible for us conceiving the broad color categories the way we do – even if we appeal to language to explain why our color concepts have the extensions that they have –, that doesn't mean that we are in any strong sense prevented from conceiving colors in any other way. In fact, the effects that were found in the empirical studies mentioned earlier are subtle, and probably reversible. If I were to learn Russian, for instance, that might lead me to distinguish more often light blue from dark blue, and I could then end up with two broad concepts rather than my current concept BLUE.[184] It is also possible that I could turn out to have these two distinct concepts by non-linguistic means, simply by creating a habit of distinguishing what are now for me two types of blue. The boundaries that concepts inherit from words are probably not absolute – they could still be expanded or diminished by other non-linguistic means. I am not saying then that we wouldn't have color concepts if it weren't for our native language, nor that, once language has causally influenced the colors concepts we have, that influence is irreversible. I just think that, as a matter of fact, the language we speak influences some of our color concepts, perhaps by forcing us to make certain distinctions that we wouldn't have had any reason to make otherwise. If we strictly follow Fodor's ideas, this influence is left unexplained.

10. The learning of color predicates

As I argued in sections 7 and 9, words play a role in concept acquisition. Just as in the case of ATOM, or of my niece's learning of ECHIDNA, it seems inadequate to say that color concepts are prior to the learning of color words, just waiting to be named. We might have thought that color concepts would be a better candidate, because colors are properties with which all people with color vision are directly acquainted. But color concepts indicate again that Fodor's view on predicate learning (the view according to which one already needs to have a concept coextensive with a predicate that is being learned) is too simplistic. But unlike the case of the concept

[184] It is less clear how learning Himba would alter my color concepts. Given that I already have the habit of distinguishing blue and green, it is unclear whether that could be reversed. I assume that I, unlike a Himba speaker, would likely have a disjunctive concept when talking about burou.

ECHIDNA, it is not as if language were just one mechanism, among many others, used to connect us to a color property. In the case of ECHIDNA, my niece could easily have acquired that concept by non-linguistic means, had she, for example, grown up in an environment full of echidnas. But living in an environment full of colors, which is where all of us with normal color vision live, wouldn't in itself grant us the same color concepts as we have now, conceived as broad categories. It seems that language directs our attention to colors, and into grouping them in certain ways, ways that are not given in colors themselves.

In learning color words, we are not simply getting connected to a color property (such as blue or green) that exists that way independently of us, like echidnas. In the case of color concepts, language is doing more than being just a link in a causal chain that connects color properties to color concepts. By teaching children color words, we inform them how they are to divide the spectrum of color, we inform them what the color properties are that we accept in our folk ontology. In hearing "this is blue", "this is green", etc., English-speaking children come to realize how they are supposed to categorize colors, that is, how they are supposed to group different hues together. It seems that, in teaching color terms to children, we are not merely giving them a means to express a concept that had, prior to the linguistic interaction, exactly the same boundaries as the words that are being learned. It is more reasonable, I think, to suppose that color concepts are being shaped by the observation of the use of color terms – even if, from a historical point of view, the terms were introduced because of cultural or individual interests or needs. Words are a means for giving our color concepts the extensions, or boundaries, that they have. In doing so, language causally determines a concept's identity, even if, from an ontological point of view, what a concept refers to can be specified with no reference to what words refer to. Had I grown up speaking Himba, or Russian, I would have different color concepts. Color names serve to show us how to categorize colors, unlike names for concrete entities, such as "echidna", which seem to either inform us that a given property (being an echidna) exists or to name a property we already conceive.

In support of the idea that color predicates are not being attached to coextensive preexisting color concepts, we can mention evidence from research in language acquisition. As Wagner et al. (2013) note, color words are words that children struggle to learn, and are only used correctly around the age of 4, much later than other words. In fact, as they note, there is a delay between a child's first uses of color words (starting typically around age 2), and its adult-like

comprehension and production of those words (around age 4). Wagner et al. experiments with English speaking children show that they make systematic mistakes in using color words. One common mistake is that of overextension of the use of a color term, especially to proximal colors. Children that consistently named a given color correctly tended to apply the same color term to other colors that are proximal to it. For instance, if a child consistently named the color red correctly, but also misapplied the word "red" to non-red things, it was more likely to apply it to proximal colors, such as orange and yellow, than to distant colors. This suggests, according to the authors, that children begin by taking color words to apply to broader categories than they in fact apply. They struggle to draw the boundaries between one color category and another, even when they already have a good notion of what good examples of certain color categories are. If they had color concepts already well delimited, we shouldn't expect this pattern of errors in their use of color words to occur, where children gradually come to apply color terms like adults.

It is interesting that Wagner et al. assume that children learn what predicates mean by forming and testing hypotheses about the extension of those predicates. According to them, "[children] learn color words by making overly broad hypotheses about their meanings, and gradually narrowing these meanings as they acquire additional, contrasting words." They also claim that "these results suggest that children begin acquiring meanings much earlier than previously supposed, and that color word learning is a gradual inductive process." (WAGNER et al., 2013, p. 313). The formulation of these hypotheses requires, as Fodor would say, representing the meanings or extensions of predicates in some way. The hypothesis must then involve the word that they are learning, and a representation of the color region to which they attribute the word. For Wagner et al., learning color terms is a mapping problem – the mapping of a word to a region of the color spectrum. But it is a mapping problem that is solved gradually.

The problem with Fodor's view, then, is that it is too simplistic, for he seems to assume that, from the start, we have concepts onto which words are mapped. As he puts it, in order to learn a language, one already needs to have a language; learning the word "green" requires having the concept GREEN. But what the case of color terms suggests is that learning color words takes time. Had children started with adult-like color concepts, the process of learning what color words refer to shouldn't take so long. So it seems that children don't come to the world knowing what concepts color words map to. They have to learn it. Throughout this process, they form hypotheses about the extensions of the color predicates, which involve the mapping of the

word to a color concept, which is a representation of a region of the spectrum to which they think the word applies. Eventually they get the right representation, as a result of linguistic training. But they start using color words before they have an adult-like understanding of the extension of the word. They get it right when they map the word to the right region of the spectrum. After a long process, they end up with a color concept, with a particular way of conceiving color categories, that is different from the concept they started the process with (which was overly broad). That is the result of learning, and in the process of hypothesizing about the extension of color predicates, they come to have color concepts that they didn't have before. In fact, the process of learning color predicates goes hand in hand with the process of learning a color concept, as I'm going to argue in section 12.

It seems, then, that the process of learning a color word can serve to shape the extensions of the color concepts we start with – precisely because the color properties are not given clear-cut in nature itself. Of course, we need to be able to perceive colors in the world, and we need to be able to perceive certain colors as more similar or more distant to certain others, in order to be able to acquire color concepts. There is likely an innate color metric that constraints the possible ways in which we can group colors together. But the precise way we end up grouping colors together, and therefore the broad color concepts we end up having, will be subject to the influence of the language we speak.

Perhaps Fodor is right that learning the word "green" requires having the concept GREEN, in the sense that the child only learns the word when it attaches the right concept to it (and this may well require that there are some innate color concepts which are later modified). But this, again, is a process that takes time, and it seems that children acquire the right color concepts as a result of linguistic training. It can still be the case that in order to understand a sentence with the word "green" in it, one needs to map the word into the concept GREEN. But from the fact that understanding a sentence in a natural language involves translating that sentence into the language of thought, it doesn't follow that in order to learn a first natural language, the symbols in the language of thought need to be already coextensive with the language being learned. So I am not claiming that the color case speaks against Fodor's view that in order to learn a language, one already needs to have a language. But if we bring to the world innate color concepts, they are not the same concepts as the ones we end up with after learning a language. So the language of thought we already need to have in order to learn a language is

unlikely to be composed out of symbols which are coextensive with the symbols in the natural language we learn. It is modified after we learn a language. Anyway, I'm simply stressing the fact that predicate learning might alter our conceptual repertoire in ways that Fodor does not seem to consider, especially in the case of color concepts, which he in many places takes to be innate. So the mapping from words to concepts doesn't work as neatly as Fodor seems to suggest.

11. Against the innateness of color concepts

I have tried to show that the empirical evidence suggests that it is unlikely that we all share the same color concepts. It seems that the way different people conceive colors is subject to the influence of the language they speak. If that is so, then Fodor was probably right in saying that words express primitive concepts – at least in the case of color words. But, to the extent that color words play a role in shaping children's color concepts, there is at least a developmental sense in which the semantics of thought is not prior to the semantics of language. Also, conceding that speakers of different languages have different color concepts clashes with Fodor's nativism.

As we've seen, Fodor was particularly committed to the innateness of sensory concepts (of which color concepts would be the typical example).[185] As he says, "we typically get the concept RED from (or, anyhow, on the occasion of) experiencing things *as red*." (1998b, p. 131). The idea that color concepts are simply triggered by instances of colored things in consequence of our biology doesn't seem right. If my color concepts are formed based on cultural and linguistic factors, then they are not innate. They were not simply triggered by color instances. Whether two hues will belong to the same color concept or not will, to some degree, depend on the culture I belong to and on the language I speak. In a way, this shouldn't be surprising. As I said, colors are a spectrum, with no obvious divisions given by nature. So it should be expected that the way different colors, of the millions we are able to perceive,[186] are going to be grouped together will admit some degree of arbitrariness. Those divisions are possibly initially drawn depending on

[185] According to him, "on all standard theories the sensorium is an innately specified function from stimuli onto primitive concepts. The range of potentially available primitive-sensory concepts is thus fixed when you specify the biological endowment of the organism, as is the range of stimulations which can occasion the availability of those concepts." (1981b, p. 276).

[186] Roberson and Hanley (2010) speak of "two million just noticeable differences".

cultural factors and interests, even if constrained by biological factors, and are then reinforced by the introduction of color names.

The color concepts we have as adults are not innate, but that is not to say that we are not born with a capacity to discriminate colors that is approximately the same in any human with three types of cone cells. And it could still be the case that we come to the world with some color concepts, or a natural propensity to group some colors together and not others. Skelton et al. (2017) present evidence that infants as young as 4 month-old group colors into categories that are actually similar to the categories commonly encoded in different languages. But they noted that infants seem to draw a categorical distinction between blue and green in a region of the spectrum that is actually taken to be focal for languages that have one composite term that applies for both blue and green, such as Himba.[187]

Something similar to what happens to phonemes could be happening to color categorization. Babies up to 10 months of age show signs of categorizing phonemes in a more fine-grained way than children and adults (see BOHN, 2000). In listening to people around them speak what is going to be their native language, they later come to be less sensitive to distinctions that are not relevant in that language. Adults can of course regain sensitivity to certain distinctions in phonemes by learning a second language and training their ears. But it is interesting that a greater sensitivity exists from early on, and that it is later lost maybe due to the lack of use. The case of color categories is not exactly analogous, for none of the experiments I mentioned showed that lacking a distinction in vocabulary yields one unable to discriminate two hues. But still, an analogy can be made in that babies seem to group colors together in a way that can be altered in consequence of experience.

Anyhow, even if we are born with a natural propensity to categorize colors in a certain way, the color concepts we end up having as adults may vary according to the language we speak, and the environment we live in. Communities that are exposed to the same colors can end up having different color words. The boundaries that my color concepts end up having are presumably not predictable from the color stimuli that I was exposed to. Since the same instances of color can belong to different color concepts for different people, my acquisition of a given color concept needs to be explained assuming the mediation of something else, not just the perception

[187] One limitation of the study is that they only tested 14 highly saturated hues. The opposition between light blue and dark blue, for instance, was not studied. Another potential problem is that the study was only performed with European infants, that typically live surrounded by artificially colored objects. So it could be that the evidence of color categorization is already a product of the environment they live in.

of colors. If cultural and linguistic factors are part of what mediates the acquisition of color concepts, then they seem to be a good counterexample to Fodor's nativism about sensory concepts.

12. Against Fodor's positive account of concept acquisition

As we have seen, in *LOT 2* Fodor does give a positive account of concept acquisition, saying that concepts can be acquired following a stage where we learn the prototype of a concept. So given that we have already shown that color concepts are not innate, and that they are influenced by color words, it is worth seeing whether this account could work for the case of color concepts. The idea would be that we first learn a color stereotype, and then, by a biological process, we lock to the color property whose stereotype we learned via language. As I've mentioned before, Margolis and Laurence (2011) formulate several challenges to Fodor's view on concept acquisition. In one criticism, they briefly consider the case of color concepts, thinking that it poses a challenge to Fodor's view because, according to them, speakers of different languages will form color concepts with different boundaries, but around the *same* stereotypes. Given that for Margolis and Laurence color concepts' stereotypes are fixed, and that for Fodor the next stage in concept acquisition is purely biological, they say that, if Fodor were right, we should all end up having the same color concepts, with the same boundaries – which they concede is not what happens.[188] In their view, Fodor's new account of concept acquisition doesn't work in the case of color concepts because it doesn't explain how our color concepts end up having the variable boundaries that they have. Given that we all learn the same prototypes, the boundaries of color concepts should all be universal, assuming with Fodor that the process of acquiring a concept is purely biological.

But, as we've seen, evidence suggests, against Margolis and Laurence, that the stereotypes of color categories, and not just their boundaries, vary across cultures and languages. So against

[188] "Focal instances of basic color concepts are highly similar across cultures; it's the breadth and boundaries of the concepts that differ. So there's nothing in Fodor's account to explain why children in different cultures end up with concepts that match their own community's way of doing things. If anything, Fodor's account predicts that children across the globe should end up with exactly the same color concepts since the same biological principles would be activated by much the same stereotypes—a prediction that unfortunately doesn't stand up to the facts." (Margolis and Laurence, 2011, p. 535).

Margolis and Laurence, Fodor's idea that concept acquisition passes through a stage of stereotype learning could be used to explain at least some of the influence that natural language has on color concepts. It could be said that natural language plays a role in our learning color prototypes, leading us to select certain hues and not others as the best examples of color categories. That way one could try to account for the variability of color concepts across speakers of different languages. And according to Wagner et al., children do learn the best examples of color categories before they learn their boundaries.

The problem is that, even if we learn different color stereotypes depending on the language we are learning, the process of getting locked to a color property, and therefore of acquiring a color concept, also involves a stage in which children learn the boundaries of each color category, as Margolis and Laurence note. And this process would hardly be explained by neurology alone.

I agree, then, with Margolis and Laurence that there is a learning process involved in the acquisition of color concepts with their respective boundaries – that this process cannot be accounted for by pure neurological processes. I am simply emphasizing that, given the variability in color prototypes, Fodor could try to place the learning process only there, and leave the rest to biology. While that would serve to account for some of the variability in color concepts across cultures, namely that of focal colors, it would still leave the variability in the boundaries of color concepts unexplained. It seems unlikely that, after learning what the prototypical red is, a purely biological process takes over, with no more need for learning. Learning the boundaries of color predicates takes a long time, and, as we've seen, some psychologists and linguists, at least, assume that this is an inductive process, in which children hypothesize about the extension of those predicates. If we assume that color concepts are the representations of the hypothesized extensions in those hypotheses, then it seems reasonable to assume that children are slowly arriving at those concepts, as a consequence of the linguistic input that they are getting from their environment. Even if it is true that children first learn the prototypes of color concepts, it seems unlikely that, afterwards, the acquisition of color concepts is simply a matter of neurological processes. The explanation of color concept acquisition won't be complete if left only to biology because there seems to be, as the linguistic evidence suggests, quite a lot of learning involved in shaping color concepts' boundaries. This is only explainable if we speak of psychological mechanisms that are subject to the influence of culture and language.

Fodor wanted to deny that concepts are learned because he thought that hypothesis formation and confirmation was a circular process. So it is likely that he would at this point interrupt and say that color concepts are not being learned because they are appearing in the hypotheses that are being formulated, and if they appear in the hypotheses, we already had them. But, as Margolis and Laurence point out, there need be no circularity in the process of formulating hypotheses about the extension of concepts. Margolis and Laurence only give examples of *complex* concepts that are learned by hypothesis formation and confirmation. But I believe the case of color concepts might be a good example of primitive concepts that can be so learned. We can assume that when children are hypothesizing about the extension of color predicates, they are simultaneously forming color concepts that are adjusting to the linguistic input they get. Experience with colors and adults using color terms can lead children to form a concept that represents a certain color property, and assume that that property is called "red". In the process of using the term wrong and being corrected, the child, after several trials, eventually gets to the concept with the right extension. So in the different stages of learning color words, children will have different color concepts associated with the words. These concepts are primitive, because they don't need to be composed out of simpler concepts. And they are likely being learned by an inductive process. They were not there from the start, but appeared and were modified as a consequence of the evidence that was available to the child at several different stages of the process. The concept that appears in a hypothesis that the child is forming in the beginning of the process is different from the concept formed later in the process. This doesn't have to be a circular process.[189]

13. Referentialism and the conventional nature of color properties

It seems, then, that if Fodor were to concede that the culture we belong to and the language we speak shape our color concepts, he would have to abandon his nativism about color concepts. I also argued that his view on predicate learning needs to be revised, and that we need to attribute a stronger role to language in color concept acquisition than that of a mediator from a

[189] I am not saying that primitive concepts can only be learned by a process that involves hypothesis formation and confirmation. My niece learned the concept ECHIDNA as a result of linguistic communication, presumably without having to formulate any hypothesis.

property to a concept. To the extent that words causally contribute to a color concept's identity, by determining the range of it referents, the semantics of language is, from a developmental perspective, causally, and in some sense explanatorily, prior to the semantics of thought, at least with respect to color concepts.

But nothing I said so far raises any problem for the idea that the content of a concept is purely referential – it can still be the case that BUROU refers to *burou*, that GREEN refers to *green*, etc., and that there is nothing more to the semantics of color concepts than reference to a color property.[190] Fodor can say that color concepts refer to specific properties in the world, regardless of what mediates their acquisition. Surely Russian, Himba and English speakers acquired different color concepts because they were raised in different cultures, speaking languages with different color terms. But what makes their color concepts mean what they do is the referential relation that they have to color properties in the world – in this case, different properties. The fact that different people have different color concepts, and that those concepts are not innate, doesn't undermine the fact that they refer to (different) properties in the world. And reference to properties is all there is to the semantics of concepts.

It is worth noting that one of the reasons Fodor wants referentialism is to preserve naturalism.[191] He wants to give an account of mental content in terms of causal relations between things in the world and symbols in the mind, with no other intentional vocabulary. As we've seen, on his view, reference is supposed to be reduced to causation. And he thinks that the mechanisms that hold the relation are not very important:

> Reference wants reliable correlations between things in the world and symbols in the head, but it doesn't care (much?) what causal mechanisms sustain such relations; in principle, the mechanisms that correlate tokens of instances of the mental representation 'C' with things in the extension of C may perfectly well differ from token to token. The trick, in naturalizing reference, is to hold onto the mind-world correlations while quantifying over whatever it is that sustains them. (FODOR, 2008, p. 143).

[190] Likewise, Fodor can still be right in saying that concept possession is purely atomistic, in that having the concept GREEN doesn't essentially involve having any other concept.

[191] He also thinks reference is the only semantic property that composes, and argues that compositionality is required to explain the systematicity and the productivity of thought, as we've seen in chapter 3.

As I argued in section 9, color properties are conventional properties. If reference to properties is supposed to be reduced to causation, then something will need to be said about how exactly conventional properties can cause anything.[192] At least in the case of color properties, we cannot simply assume that the means by which we come to refer to these properties is irrelevant. The visual perception of colors, in and of itself, doesn't give us the boundaries in the color spectrum. For, after all, what we perceive are particular hues. It is not true that "all that's required for us to get locked to *redness* is that red things should reliably seem to us as they do, in fact, reliably seem to the visually unimpaired" (1998a, p. 142). If that were all that was required for us to get locked to a color property, it seems that we would all end up dividing the color spectrum into the same color properties, provided we had contact with the same colors. But that is not the case. So more is in fact involved in our getting locked to, say, blueness, or goluboyness.

In the case of color concepts, it is reasonable to say that the language we learned played a role in determining which color properties we were going to get locked to. The cultural and linguistic mediation that gets us from instances of a color property to a color concept is part of what institutes a color property for us. Mere causation from particular colors to a concept, abstracted from the mechanisms of mediation (which are intentional), won't work to explain how we got locked to a given color property in the first place. Since color properties are conventional properties, as such, they require some extra mediation, which is not merely biological, but intentional in nature, in order to cause a concept in us. The linguistic mediation we have is what explains why we take certain color properties and not others to belong to our folk ontology, and how we came to refer to them. Something that supports the idea that color properties are conventional is the fact that children struggle to infer the boundaries of the color categories, as Wagner et al. (2013) point out, which suggests that the limits of color properties are not clearly given in nature.

So I am here simply reinforcing the idea that language plays a causal role in the *acquisition* of our color concepts – a stronger role than Fodor attributes to it. But again, that is compatible with the view that what *constitutes* the *content* of a concept, what plays a role in individuating the concept, is the property that it refers to.[193] And that may be independent of language. Getting

---

[192] The same problem applies, more obviously, to properties such as *king*, *Tuesday* and *money*. Fodor addresses this issue in *Concepts*.

[193] "If, for example, DOORKNOB is primitive, then whatever metaphysical story we tell about the *content* of primitive concepts has to work for DOORKNOB. And so must whatever psychological story we tell about the

locked to a color property usually involves a stage where language learning causes us to select certain colors as members of the same category, but that doesn't mean that language is constitutive of color concepts identities. It doesn't even mean that language is a necessary mediation, or required, for my acquisition of a concept that refers to blue things, for instance. We lock to blueness via the English language de facto, but not metaphysically. It could have been otherwise.[194] So this view of the influence of language on thought says nothing against Fodor's informational semantics, according to which "content is constituted by some sort of nomic, mind–world relation. Correspondingly, having a concept (concept possession) is constituted, at least in part, by being in some sort of nomic, mind–world relation." (1998a, p. 121) After language plays its part in concept acquisition, it doesn't play any part in what constitutes a concept. The concept is locked to a property, and even though, in the case of colors, that happened because of the particular language I learned, the role of language was causal, and not constitutive of what that concept is.

A related worry that could come up here is the following: if thought is explanatorily prior to language and therefore linguistic meaning is to be explained in terms of thought's content, then we should not expect the content of thought to be explained in terms of linguistic meaning. Saying that the content of thought is to be explained in terms of causal relations with properties in the world would then be a way of saying that it is naturalizable, and that it is independent of the content of language. But in saying that my concept BLUE means what it does in part because of causal relations to the world that were mediated by the word "blue", are we not explaining the content of thought in part by appeal to linguistic meaning, rendering the explanation circular? I think that a proponent of the priority of thought could insist on the idea that the mediation of language is not essential to my concept having the content that it has. While the meaning of a word will always be explained in terms of the concepts that it is used to express, when a concept first appears, its content will always be explained by causal relations which are never mediated by a word with the same meaning. A word can then be introduced and be a link in the causal chain that eventuates in others acquiring the same concept. But its mediation is never essential.

---

acquisition of primitive concepts. And the metaphysical story has to work in light of the acquisition story, and the acquisition story has to work in light of the metaphysical story." (1998a, p. 123)

[194] Language, along with perception, may be sufficient to give us the color concepts that we have, but I wouldn't say it is necessary. It could be that extensive training in grouping certain colors together, as a separate group from others, could be something achieved by non-linguistic means, and it could conceivably yield to the same color concepts that we have as a result of speaking a given language.

For all we've seen, then, Fodor could still hold that reference reduces to causation.[195] I just wanted to highlight that the initial causation in question, which eventuates in our acquisition of a concept, will not always go neatly from instances of a property in the world to a mental symbol. It will often be heavily mediated by language. It is worth adding that, if the color properties that we refer to are determined conventionally, then, against Fodor, the contents of at least some thoughts are determined conventionally. Fodor is not completely right in saying that "semantics is essentially a theory of thoughts, the contents of which are, I suppose, *not* determined by norms and conventions. Quite possibly English has no semantics, some appearances to the contrary notwithstanding." (2008, pp. 198-9).

14. Color terms and perception

In the beginning of the chapter I presented some arguments against the idea that the vehicle of thought is a natural language. We've seen, for instance, that the ambiguity of natural languages speaks against their role as the vehicle of thought. I then tried to show that the particular language we speak can shape at least some of our concepts, such as color concepts. These two views may seem incompatible, but they are not: even if language is not the vehicle of thought, it can still shape some of our concepts. In suggesting that the language we speak can shape our broad color concepts, in the sense of causing them to have the extensions that they have, I don't mean to be committed to the idea that natural language has to be the vehicle of color thoughts, or that it is necessarily involved in those thoughts, to use Carruthers' terminology.

It is worth noting that some studies on the categorical perception of colors have suggested that linguistic representations are actively involved in the perception of colors. The idea is that language plays an online role in perception. In the study I mentioned earlier by Winawer et al., involving Russian and English speakers, subjects were shown three color squares, and had to indicate, as quickly and accurately as they could, which of the two bottom squares was identical to the top square. But in order to test whether language had an online influence on color

[195] Another problem in this case is that, to the extent that properties are abstract objects, it could still be argued that properties never cause anything, only their instances do. Perception, for instance, only gives me particular shades of blue; it doesn't give me blueness. So how do I come to refer to the property of blueness, which includes in fact several different hues? This seems to be a problem for the idea that reference reduces to causation from things in the world to symbols in the mind for, strictly speaking, there is no blueness, or goluboyness in the world.

discrimination, subjects performed the test under three conditions: one with no interference, one in which they had to simultaneously perform a verbal task, and one where they had to simultaneously perform a spatial task.[196] What they found was that Russian speakers showed a category advantage in trials without interference, and with spatial interference, but not with verbal interference. From this they claimed to have shown that language plays a role in the visual discrimination of color, for the category advantage that Russian speakers exhibited in normal trials was eliminated in trials with a verbal interference. According to them, category advantage disappeared in trials with verbal interference because linguistic representations, that normally would be acting on perception, were being recruited for something else. As they put it,

> our results suggest that language-specific distortions in perceptual performance arise as a function of the interaction of lower-level perceptual processing and higher-level knowledge systems (e.g., language) online, in the process of arriving at perceptual decisions. (…) it appears that language-specific categorical representations play an online role in simple perceptual tasks that one would tend to think of as being primarily sensory. (WINAWER et al., 2007, p. 7784) [197]

There have been other studies that supposedly showed that language is actively involved in categorical perception effects. In a study by Gilbert et al. (2006), English speakers were instructed to fixate on a cross that was surrounded by twelve colored squares. The color of one square, the target, differed slightly from the color of the other squares. Subjects had to indicate as quickly and accurately as they could in which side of the circle, left or right, was the target located. The colors used in the experiments were two shades of blue and two shades of green. In some trials, the target belonged to the same category as the other squares, being just a different shade of the same color, while in other trials the target belonged to a different category (e.g. the

[196] In the trials with verbal interference, "subjects were given an eight-digit number series to rehearse during the color task. This series was presented for 3 sec, and subjects were instructed to rehearse it silently." (p. 7785) In the spatial interference conditions, "subjects viewed a 4 X 4 square grid of which four random squares were shaded black. Subjects were instructed to remember the grid pattern by maintaining a picture of it in their mind until tested." (p. 7785) Even though the authors pretested the difficulty of each of these tasks, finding that subjects had the same level of accuracy in both, it is still possible that both tasks would not be equally difficult when combined with another one in a dual task experiment. So it is possible that performing the experiment with a verbal interference was a more demanding job than performing it with a spatial interference.

[197] "The fact that Russian speakers show a category advantage across this color boundary (both under normal viewing conditions without interference and despite spatial interference) suggests that language-specific categorical representations are normally brought online in perceptual decisions." (pp. 7783-4) According to them, "These results demonstrate that categories in language can affect performance of basic perceptual color discrimination tasks. Further, they show that the effect of language is online, because it is disrupted by verbal interference. Finally, they show that color discrimination performance differs across language groups as a function of what perceptual distinctions are habitually made in a particular language." (p. 7783)

target was blue and the other squares were a close shade of green). They found that categorical perception only occurred when the target belonged to a different category from the other squares and was presented in the right visual field (RVF), but never when the target was presented in the left visual field (LVF). That is, the time it took participants to indicate that the target was on the left side was approximately the same, whether or not the category to which the target belonged was the same or different from the color category of the other squares – there was no categorical perception effect on the LVF. But when the target was presented in the right side, participants were faster to identify it when it belonged to a different category than when it belonged to the same color category as the other squares.

Given that the brain processes the information from each visual field in the opposite cerebral hemisphere, and that the left hemisphere is the one thought to be predominantly involved in language processing, they explain the finding by saying that targets that appear in the RVF are processed faster by the left cerebral hemisphere, where linguistic representations can interfere. This means, according to them, that linguistic categories are being accessed and are interfering in discrimination when a target is presented to the RVF, but not when it is presented to the LVF.[198,199] They also found that the categorical perception in the RVF disappeared when subjects had to perform the task with a verbal interference. They claim that their results

> are consistent with the hypothesis that linguistic categories selectively influence color discrimination in the RVF. Color names modulated color discrimination, enhancing between-category distinctions and perhaps reducing within-category distinctions, but only when the target appeared in the RVF. These effects were disrupted by verbal interference. (GILBERT et al., p. 490).

But the idea that names, or verbal labels are interfering in perception is challenged by Holmes and Wolff (2012). In their experiments, they observed categorical perception in the left

---

[198] As they observe, "given the contralateral nature of visual projections to the cortex, information from the right visual field (RVF) would, at least initially, have preferential access to, and be more susceptible to modulation by, lexical representations in the LH. This fact suggests that an effect of language on perceptual discrimination would be seen primarily for stimuli in the RVF." (GILBERT et al., 2006, p. 489).

[199] Also highlighting the online role of language in perception, Roberson et al. note that "these findings raise the possibility that CP [categorical perception] for color is mediated by higher-level cognitive processes rather than by perceptual warping, and only occurs when a linguistic code is accessed. According to this view, CP occurs because colored stimuli from different color categories are represented as separate terms (e.g. blue vs. green) in a verbal code. Consequently, they can be distinguished rapidly because both verbal and perceptual codes provide converging evidence that they are different. Two different shades of the same color category will be distinguished more slowly because they will activate the same verbal label, which will conflict with the perceptual information that they are different." (ROBERSON et al., 2009, p. 483).

hemisphere both with labeled and unlabeled categories. They suggested then that there is no evidence that lateralized categorical perception is driven by language, or by linguistic representations. According to them, the left hemisphere may be responsible to partition experience into categories, whether or not they have a name. As for the elimination of the categorical effect with verbal interference, they note that categorical perception "may be disrupted more by verbal than by spatial interference (e.g., Winawer et al., 2007) because the former disproportionately taxes not only linguistic processing but also left hemisphere processing independent of language" (2012, p. 442). Their findings suggest, I believe, that color names are not necessary for the occurrence of categorical perception, for it can equally well happen when we discriminate between things whose categories don't have a name. As they put it, "while it remains possible that CP might sometimes be driven by language, the more parsimonious conclusion is that CP, even for labeled categories, is driven by nonlinguistic factors." (p. 442). Color concepts or categories, which are nonlinguistic, might then be what explains categorical perception.

Another reason for not assuming that linguistic representations are responsible for categorical perception effects is that, as I mentioned, some studies have found evidence of categorical perception of color in pre-linguistic infants as young as 4 month-old (see Skelton et al., 2017). This is another type of evidence for the idea that the categorical perception of color doesn't require language, given that even very young babies seem to group colors in categories. Instead of assuming that language, once learned, takes over and is responsible for categorical effects on perception and memory, it would again be simpler to assume that color concepts or categories are what explain the partitioning of colors in categories from the start.

Concluding that linguistic labels, or words, are interfering in color perception, merely based on the fact that CP is disrupted by verbal interference, or that it only happens when a target is presented in the RVF, is too rushed a conclusion. For it is not as if the left hemisphere is solely dedicated to language. For all we know, what can be interfering in perception is concepts, or categories, but not necessarily words. So instead of assuming that color names are modulating perception, it seems more plausible to assume that language influences concepts, and concepts act on perception, and not that words are actively involved in perception. This is compatible with the view that language is an important means (though not the only one) we use to form categories, or to reshape preexisting ones, even if words are not involved in the subsequent use of those

categories. Unlike what most researchers working on linguistic relativism suggest, we don't need to assume that natural language words are the vehicle of color concepts to account for the categorical effects that have so far been observed. Words can influence some of the concepts or categories that we have, but they can still be conceived as dissociated from them – as not typically involved in color perception, memory and thought. That is, I think, a more plausible interpretation of the findings.[200]

15. Conclusion

My account so far may seem to have been mainly negative. Using the case of color concepts, I have tried to show where Fodor's views on concepts went wrong. But in doing so, I think some positive views inevitably follow. Making explicit the problems of a view can take us closer to seeing what is right. So I don't intend what I said to be purely negative. Some positive morals can be drawn. For instance, in arguing that color concepts are not innate, I'm making a case for the thesis that they are learned. In arguing that color predicates are not merely naming preexisting coextensive concepts, I'm strengthening the idea that language and culture are causally responsible for giving some of our concepts the extensions that they have. In saying that natural language's function is not only that of communicating thoughts, I'm saying that it also serves other functions, such as that of establishing conventions on how to categorize certain things, which in turn bias us into conceiving things in certain ways and not others.

I want to stress here that I'm not trying to defend a form of linguistic relativity according to which all our concepts are completely created and shaped by the language we speak, as if there was no prior organization in our minds before the acquisition of a natural language. Among other

[200] Are these categories that are affecting perception the same as the ones we use in thought? I don't know. I'm assuming here that it is at least reasonable to assume that differences in performance in memory and perceptual tasks among speakers of different languages suggest that these speakers don't share the same color concepts. But it could be that, in the case of Roberson et al. memory tasks, participants were relying on color names of their language to recall stimuli. And in the case of studies involving color perception, something other than concepts could be affecting perception. The habitual distinction that they make of such colors, in order to speak, may have altered some perceptual category, that is not necessarily the same as a concept that is accessible to thought and reasoning. If that is the case, them the studies mentioned so far would only be evidence for the influence of language on memory and perception, but not on thought properly. I think, though, that one way to investigate whether language influences thought would be to simply ask speakers of different languages what color properties they think are there in the world. Assuming that English speakers would say that there is blue, green, etc., that Himba speakers would say that there is burou, etc., that Russian speakers would say that there is goluboy, siniy, etc., this would be evidence of what they take to be the color properties that there are – assuming that they are using language to express what they think.

things, it would be hard to explain how non-linguistic creatures could survive with a mind incapable of somehow organizing experience. I am not even trying to suggest that color concepts are generated by color terms, or that they would not exist otherwise. It is reasonable to assume that continuous experience with colors could give someone color concepts even without the use of language. Also, the variation in the color vocabulary across different languages forces us to accept that the human mind is capable of forming color concepts in a variety of different ways. Our minds seem to be very flexible. Were we not capable of thinking about colors in a variety of different ways, there would be no difference in color vocabulary across languages. It would also be impossible for a speaker of English to understand that speakers of Himba categorize colors in a different way, were English speakers unable to conceive other ways of categorizing colors. So it seems certainly right that some thought needs to exist before language, and that, after learning a language, we can still think about things in ways that don't strictly conform to our native language.

I generally agree with Pinker (2007), who says that proponents of linguistic relativity tend to reverse cause and effect, assigning a greater influence of language on thought than sometimes is reasonable (like in the case of snow). Relativists generally neglect the relevance of non-linguistic factors in cognition, emphasizing the role of language. But Pinker and Fodor, I think, seem to move to the other extreme, attributing little importance to the influence words can have on our conceptual scheme.

It is certainly an empirical question which concepts, if any, are shaped by the language we speak. My suggestion is that at least when it comes to concepts that express non-discrete properties, where there is greater arbitrariness in how the world is to be divided (of which color concepts are a paradigmatic example), words can play the role of creating uniform concepts among different people. Without attributing language a function other than that of merely expressing thoughts, it would be a mystery how some groups of people end up with approximately the same color concepts, given our inborn capacity of categorizing colors in a variety of ways. Color words allow members of the same community to be, as it were, on the same page about colors. Were they each left on their own to divide the color spectrum at will, we would expect to find across individuals the same conceptual variation that we find across cultures. Properties that are not clear-cut can potentially cause concepts with different boundaries in different people. So language is a means of expressing thoughts, but it is also a means of imposing

boundaries to the concepts of others, concepts that would otherwise be too variable across individuals. I do think that the main function of language is communication, but in order for communication to work, we need to have some agreement about what we are talking about, and how we are going to classify things. So we also use language to make some of our concepts shared by other members of our community.

It is important to note, though, that the fact that natural language can influence thought, or the way we conceive the world, says nothing against the idea that we think in a language of thought that is different from any natural language. Unlike what Fodor at times assumes, the language of thought doesn't have to be universal or innate. The way we conceive and categorize properties in the world will be subject to the influence of several factors, including the language we speak. But still, from the fact that there is variation in color concepts across cultures which correlates with linguistic variation, it doesn't follow that we think in a natural language. This point has certainly been overlooked by sympathizers of linguistic relativism. Nicholas Evans, for instance, says the following:

> generative linguists and philosophers, from Jerry Fodor to Stephen Pinker, (…) postulate that humans think in a language-independent "mentalese," which is then translated straightforwardly into English, Japanese, Dalabon, or Navajo. (…) [T]he question to ask is: what exactly are the concepts that are used in "mentalese"? Is the "give" in mentalese the English "give" or the Navajo one, is the "leg" the English or the Savosavo one, is the "think" the English or the Dalabon one? Only a vanishingly small group of entities are carved at the same joints across different languages (…). Once we take a broad sample of the world's tongues into account, it is clear that even quite basic concepts just do not line up across languages. Their semantic diversity suggests that Edward Sapir was closer to the mark than Aquinas, Fodor, or Pinker. (EVANS, 2010, p. 59).

But the fact that language influences conceptualization says nothing against the idea that we think in Mentalese. A proponent of Mentalese doesn't have to be committed to the idea that thought is completely free from the influence of a natural language. And even if language does influence some of our concepts – in a way that speakers of different languages may not share all their concepts – that doesn't mean that no concepts are universally shared. Some concepts may be universally shared, even if they aren't always expressed by monomorphemic words in every language. For instance, in the case of color concepts, I am assuming that language can influence the general way we conceive colors, or the way we group colors together. But I can still share

more specific color concepts with speakers of different languages – a Russian speaker and me can both have the plan to paint our houses the same color as the color of the sky on a clear day. We have more concepts than we have words, and some of these concepts will be shared, even if they are not always expressed linguistically in the same way. Besides, I take it that the arguments previously presented, against the idea that we think in a natural language, still hold. So the vehicle of thought may still be a language different from any natural language, even if natural languages influence some of our concepts.

I'm suggesting then a view that is somewhat a mix of a communicative conception with a cognitive conception of language – but not cognitive in Carruthers' sense, in which language is assumed to have a cognitive function only if it is itself used in thinking, as a vehicle for thought. This mixed view goes approximately like this: we come to the world prone to form certain concepts about things that are most salient in our environment and relevant for our needs – much like what happens to most animals. We humans perhaps have a greater flexibility in our capacities for attention and categorization, which allows us to form concepts about a wider range of things than, say, cats and dogs. In developing language, we started to put labels on things for which we perhaps had well defined concepts, but we also started labeling things or properties whose boundaries are not so clear in nature. In labeling those properties, we are likely guided by similarities that we recognize in nature, by biological propensities, and by ecological salience. But still, since by assumption some properties are not clearly divided in nature, in naming them we tacitly agree on what is approximately to count as a member of a given category. In teaching that word to others, we guide them to conceive the world in a similar way – a way that was not necessarily given to them before. So we use language not only to express our thoughts, and to transmit beliefs, but also, as Whorf envisioned, to carve up nature, to shape our concepts in a way that puts everyone in a linguistic community on the same page, as it were, about how they are going to categorize certain features of the world, especially those whose division is not clearly given by the world.

I argued that that is precisely what happens in the case of colors. The way we group colors together into categories (red, yellow, blue, green, etc.) is not given clear-cut in nature, and color words have the function of shaping our color concepts (even if the possible variation in color categorization is constrained by the structure of color perception, which is biologically determined, and by colors that are present in the environment). In learning the color words of our

language, we are trained to make certain distinctions and we develop a habit of thinking of colors as pertaining to a certain group and not to others (even if the colors in the boundary region won't be so obviously members of a determinate category). Words guide some of our habits – in this case, the habit of distinguishing certain colors.[201]

But even if language is one way of shaping the boundaries of concepts, it likely isn't the only one. Habit, or experience in general could suffice in some situations. A possibility is that language is just another way of creating habits, though a mandatory one (for anyone who uses language), which forces our attention in certain directions. But a radiologist will be good at noticing differences in an x-ray, a botanist will be good at noticing different trees, and so on. All those habits, things we do frequently, will have the power to change the extension of concepts, to change the way we separate and group things together in the world.

So the language we speak may bias us to categorize certain properties in a certain way, because, in speaking the language we do, we get accustomed to categorizing things in the way our language requires us to. In some cases, we can come to group certain things together that don't, in any necessary way, belong together. But, to use Carruthers' distinction, I'm claiming neither that language is required (in a strong sense) for us to have color concepts nor that language is constitutively involved in color thoughts. Language may be required for us to have thoughts about genes, but it does not seem required for us to have color concepts. In practice, it allows us to have color concepts that are shared by people in our community. But it is conceivable that we could have achieved this by other, non-linguistic means. And even if language does determine the extension of some of our color concepts, I see no reason to assume that language is involved, or that it is the vehicle of the color concepts that it shapes.

To sum up, Carruthers' opposition between a communicative conception of language and a cognitive conception, introduced in section 2, doesn't clearly accommodate the role that I

[201] I have assumed in this chapter the concepts such as ECHIDNA are not shaped by language, though they are typically acquired through language. But one could argue that even in respect to discrete entities, such as echidnas, language does play a role in establishing a consensus about what is to count as an entity of that kind. For instance, a person living in Australia might see echidnas and porcupines around and think that they are the same animal. Through linguistic communication, one can inform that person that they are actually different animals, which would in turn modify that person's concepts. So in a situation like this, it is possible that language would be shaping our concepts, just like in the case of color concepts. If that is so, then the distinction I'm establishing, between the role of language in shaping concepts for discrete and its role in shaping concepts for continuous entities is perhaps of no great significance. However our imaginary Australian could have come to find out that they are different animals though non linguistic means, whereas colors are conventional properties, so there is no such possible discovery.

attributed to language here: that of shaping some of our color concepts, perhaps as a consequence of us using the color words in our language and making the distinctions that they force us to make.[202] This view is somewhat in between Carruthers' cognitive conception of language, and Fodor's communicative conception. As for Fodor's views discussed here, I think that, at least in the case of colors (and presumably of other properties that are not clearly divided in the world), we can stick to Fodor's idea that monomorphemic words express primitive concepts. General color concepts such as BLUE, or BUROU, cannot be broken down into simpler concepts, out of which they are composed – though they may be structured around best examples, or prototypes. We can also preserve the idea that color concepts are at least in part individuated by their extensions – this is what leads us to say that speakers of different languages have different color concepts. But the ideas that color concepts are innate, that learning color predicates is simply a matter of naming preexisting color concepts, that reference reduces nicely to causation from color properties to color concepts, and that the semantics of thought is prior to the semantics of language, have to be revised or qualified, for they are either incompatible with or too simplistic in the face of the variation in color terms, and concepts, that is found across different cultures.

[202] In fact, perhaps the whole scheme of classification Carruthers proposes should be revised, and the opposition shouldn't be drawn only between those who think that we use language to communicate thoughts, and those who think that natural language is involved in at least some of our thoughts. It is also relevant to consider the degree to which one thinks language can determine the identity of (some or all) the concepts we have. An extreme linguistic relativist could maintain that all the concepts we have are shaped by language, or acquired though language (to a point where we wouldn't have concepts if we had no natural language) while still maintaining that thought is always different from language – that no thought involves language. This idea, that language is necessary for us to have concepts, is distinct from the idea that we use language to think – that is, it is compatible with the view that we think in a language of thought, even if this language of thought is supposed to be structurally and semantically indistinguishable from the natural language we learned. But, according to Carruthers' classification, this view would still be part of the communicative conception of language, which is a misleading way of classifying it.

CONCLUSION

In this dissertation, I have tackled different topics related to the language of thought hypothesis. In the first chapter, I introduced the representational and the computational theories of mind. We have seen that the representational theory of mind treats propositional attitudes as relations that we have with mental representations, which, according to the language of thought hypothesis, have a linguistic structure. The computational theory of mind is a theory about mental processes. It says that (at least some) mental processes can be treated as computational processes, which are understood as syntactic transformations of symbols in the language of thought. I also tried to clarify the extent to which Fodor can be considered a functionalist. Roughly, he thinks that beliefs have a particular functional (or causal) role in a system. Desires, on the other hand, have a different functional role. The exact differences in functional role of the different attitudes are to be investigated by psychology. But the belief that *P*, or the desire that *Q*, are not states that are to be individuated by their functional role. If mental states with semantic content were so individuated, that would lead to holism, and to the falsity of common sense psychological explanations – which assume that people share beliefs and desires. Finally, I tried to make clear the scope of RTM and CTM. Someone might object that they don't offer a complete naturalistic vindication of common sense psychology. RTM doesn't explain, for instance, mental states, like sensations, that have a qualitative character, nor the opposition between conscious and unconscious propositional attitudes. But, in fact, RTM is not intended to be a complete theory of the mind. The hope is only that it explains propositional attitudes. Assuming it does that satisfactorily, there is no need to demand anything extra from it. After all, it is not as if we have other competing theories that explain both conscious and intentional mental states. I also indicated that CTM is somewhat limited in its scope. It doesn't explain, for instance, associative processes. However, Fodor makes it very clear that its goal is to explain only rational, inferential processes. But even here, I have tried to show that, as Fodor conceives CTM, it only explains modular or demonstrative processes, that is, processes that involve local syntactic properties of representations. It doesn't explain everything there is to explain about rational processes, such as abductive inference.

Still, there are some advantages in adopting RTM and CTM, as opposed to reductionist theories about the nature of mental states, such as behaviorism and the identity theory. In chapter 2, I tried to show that RTM offers a better vindication of intentional common sense psychology, for it, unlike behaviorism and the identity theory, preserves the two properties we ordinarily attribute to propositional attitudes: causal efficacy and semantic evaluability. Another general theme of this chapter was the relation between the psychological level of explanation and the level of explanation of cognition that belongs to neuroscience. In part 2, I discussed the multiple realizability argument, and how it relates to psychology's autonomy from more basic sciences. The multiple realizability argument, as we've seen, is typically used to falsify the identity theory, and to preserve the autonomy of psychology. I argued that multiple realizability, though a very plausible assumption (in particular when it comes to propositional attitudes) is an empirical hypothesis. If true, it is sufficient to deny the identity theory and to establish the ontological autonomy of psychology with respect to neuroscience. Though perhaps sufficient, multiple realizability doesn't seem necessary to deny the identity theory, nor to establish the autonomy of psychological explanations (at least from an epistemological perspective). In the final part of chapter 2, I considered some of Searle's arguments for the view that computational processes are not sufficient, or even necessary, to explain intentionality. I tried to imagine an exchange between Searle and Fodor, suggesting, in the end, that an intermediate position is more plausible. Roughly, I think Searle is mistaken to think that a computational description of the mind, including the language of thought, plays no explanatory role in cognition. But, against Fodor, it may well be that intentionality is a product of creatures with brains like ours.

In chapter 3, I discussed the main arguments for the language of thought. So even if it is true that only some mental processes can be treated as computational processes, that doesn't mean that there are no other reasons, besides CTM, to assume that there is a language of thought. There I introduced first the productivity, then the systematicity argument for the language of thought. The idea is that thought seems to be productive, because we are constantly having thoughts we never had before. Thought also seems to be systematic, because the ability to think some thoughts is related to the ability to think other related thoughts. These properties are explained by the assumption that the vehicle of thought is sentences in a mental language. It could still be said, though, that connectionist models present an alternative to the assumption that there is a linguistic system of mental representations. This is why, in section 3, I briefly present

some of Fodor, Pylyshyn and McLaughlin criticisms against connectionist models of the mind. Their view is that connectionist models don't explain the systematicity of thought, but they can still be conceived as implementation models of the language of thought, instead of competing cognitive models of the mind. Finally, in section 5 I discuss Fodor's argument against the compositionality of language, which is intended to show the priority of the content of thought over the content of language. There I argue that Fodor's argument is circular, and that it creates problems for his productivity and systematicity arguments for the language of thought.

Assuming we are convinced by the productivity and systematicity arguments, and that we accept that we think in a language of thought, we might not yet be convinced, one way or another, as to whether the language of thought is a natural language. In chapter 4, I turned to the issue of how the language of thought is supposed to be related to the natural languages. I began by introducing some reasons to think that thought is independent from natural language, and therefore that the language of thought is not to be identified with any particular natural language. I also discussed some of Carruthers' ideas, such as his opposition between a communicative conception of natural languages (according to which the role of language is to express thought) and a cognitive conception (according to which natural languages are the vehicle of at least some thoughts). We have seen that Carruthers holds a version of the cognitive conception of language, according to which our conscious propositional thoughts have natural language sentences as their vehicle. Fodor, on the other hand, holds a communicative conception of language. He also thinks that words typically express primitive concepts, that most primitive concepts are innate, that the only semantic property concepts have is reference and that reference reduces to causation, and that the semantics of thought is prior to the semantics of language. I then argued that probably not all of these views can be true, if we accept some empirical findings about the influence of natural language on our conceptualization of colors. I tried to show how this creates problems for several of Fodor's views. In the end, I argued for an intermediate position about the role of language, one that is not clearly contemplated by Carruthers' opposition between communicative and cognitive conceptions of language. I argued that natural languages can be causally responsible for determining the extension of at least our general color concepts, but that that doesn't mean that language is the vehicle of the concepts whose extensions it shapes.

A note on H_2O and XYZ

There are several other questions related to the language of thought hypothesis that were barely, if at all, touched in this dissertation. I didn't discuss in any detail, for instance, the issue of the semantics of the primitive symbols (concepts) in the language of thought, and more generally the issue of concept individuation. This would have involved dealing with questions such as: when do different people, or the same person at different times, have the same thoughts and concepts? What makes a concept be the concept that it is? I will end with a brief note on that.

One traditional way to approach these issues is via the internalism/externalism debate. Roughly, internalists hold that the content of a concept supervenes on the intrinsic (non-relational) properties of an organism – so two physical duplicates would share all their concepts. Externalists, on the other hand, hold that the content of a concept can be individuated by relational properties – so two physical duplicates, in different environments, might not share all their concepts. The classic thought experiment here is the one offered by Putnam, of Oscar1, who lives on Earth, and his twin, Oscar2, who lives on twin Earth. The only difference between Earth and twin Earth is that the liquid we here call "water" is, on twin Earth, composed of XYZ, instead of H_2O. What Oscar2 calls "water" (in twin English) shares all its macroscopic and functional properties with what Oscar1 calls "water", but its chemical composition is completely different.

Let us assume that Oscar1 and Oscar2 have the same behavioral dispositions towards that liquid, and thought processes that are indistinguishable by their causal roles. They also have exactly the same brain states. Should we say that they both have the same WATER concept? Or should we say that their concepts are different, because they are about different substances? An internalist would say here that, despite the differences in the chemical composition of water, Oscar and twin Oscar share the concept WATER, for after all there is no difference in their brain states, and contents are supposed to supervene on intrinsic properties. Besides, there is no difference in their behavior, or in what they believe about that liquid (let's assume they don't know anything about chemistry). The externalist would say that their concepts are different, because H_2O and XYZ are different substances, and therefore their thoughts are about different things (they have different truth-conditions), and thoughts should be individuated, at least in part, by relational properties, or by whatever they are about.

In *Psychosemantics*, Fodor argues that sciences, including psychology, individuate entities by their causal powers. Oscar1 and Oscar2's mental states have the same causal powers, given that their behaviors are indistinguishable. Their mental states differ in their relational properties (one relates to H_2O, the other to XYZ), but these differences are not relevant for psychology, because they don't affect the causal powers of their mental states. So for the purposes of psychological explanation, Oscar1 and Oscar2 share the same mental states and concepts. Psychology is internalist. From a psychological point of view, there is no reason to distinguish the twins' mental states.[203]

In later work, Fodor goes on to argue that concepts should be individuated by their syntax and reference. The view that reference is part of what individuates concepts relates naturally to externalism. This way of individuating concepts seems to imply that twins have different concepts, because Oscar1's concept WATER refers to H_2O, whereas Oscar2's concept WATER refers to XYZ. So in saying that concepts are individuated in part by their referents, it seems that we need to take into account the relational properties of Oscar1 and Oscar2's mental states, and to distinguish their WATER concepts, despite their having identical causal powers.

Is there a way of preserving both the idea that concepts supervene on non-relational properties (which is what psychology seems to require) and the idea that reference is part of what individuates concepts (for this particular case, at least)? I believe so, if we accept that there are levels of reference, and that we might not need to go deep into the microstructure of entities to determine what the referent of our concept WATER is. So even if we accept internalism, there is no need to deny that what happens in the environment is relevant to making our concepts the concepts that they are. Accepting that the environment is relevant to determining the identity of our concepts would not force us to say that Oscar1 and Oscar2 have different concepts, for we might say that Oscar1 and Oscar2 are in fact acquainted with the same *perceptual* environment. Whatever goes on in the microstructure of water here and on twin Earth is of no concern to them, for H_2O and XYZ present themselves in the same way to both of them. So in some sense of "being about", Oscar1 and Oscar2's thoughts are about the same thing – that transparent liquid that runs in the rivers, falls from the sky and quenches thirst. Both Oscar1's and Oscar2's concept

[203] In fact, in *Psychosemantics*, Fodor thinks that what the twins share is a "narrow content", conceived as a function from context onto truth-conditions. So when Oscar1 and Oscar2 think that water is wet, they share the same function, but given that they are in different contexts (one where there is H_2O and the other where there is XYZ), their thoughts have different truth-conditions, so their "broad content" is different (Oscar1's thought is true iff H_2O is wet whereas Oscar2's thought is true iff XYZ is wet).

WATER refer to the same thing – a liquid that has some functional properties such as being drinkable, and that presents itself perceptually in a certain way. But to the extent that they are oblivious to the chemical structure of that liquid, the chemical level of reference is not relevant for the content of their concept, from a psychological perspective. It doesn't change their concepts' content or bcausal powers.

But how about after the invention of chemistry? Does learning that water is composed of hydrogen and oxygen change our concept WATER? I believe that coming to believe that water is H_2O, in fact, doesn't essentially change our concept WATER. The concept WATER still has as its referent the substance in the world that is transparent, eliminates my thirst, etc. What changes is that I acquired a new belief about water, a belief about its microscopic structure that I didn't have before. I acquired a new concept, namely H_2O, which captures whatever goes on in water at the chemical level of reality. My twin will have a different belief about water than I do – she will think that water is XYZ and not H_2O. But her belief is still about water, that is, it is still about water as water presents itself to organisms with the perceptual systems that humans have, and which has the functional property of being what eliminating the thirst of most animals. So nothing essentially changed in the concept WATER when we found out what molecules water is made of, because what is characteristic of that concept is that it captures a property at a macroscopic or perceptual level of reality. We simply acquired a new belief about water, one which says that the substance we call "water" is constituted, at the chemical level, by molecules which have two atoms of hydrogen and one of oxygen. I think WATER, and I can think H_2O, but the two are still different concepts, because they have as referents properties in two distinct levels of reality: one is about the substance that plays a fundamental role in our lives, and the other is about its chemical constitution.

My point, then, is that the idea that reference individuates concepts can be accommodated to the idea that concepts supervene on intrinsic properties of organisms – at least in the case of water and twin-water. All we need is to admit levels of reference. Besides, what matters to psychology is what people are aware about the entities they think about. If we want to preserve the idea that the ordinary person usually knows what they are thinking and talking about, when they talk about water, we don't have to abandon referentialism. We only need to distinguish between levels of reference, and say that what the ordinary person thinks about when they think about water is that substance which is given in their perceptual circle (to use Fodor and

Pylyshyn's (2015) expression) – regardless of whether they know or not that that substance is made of H_2O. We shouldn't have to do chemistry in order to find out what people's thoughts are about – not in a sense of being about that is relevant for psychological explanations. But these ideas are to be developed in the future.

REFERENCES

AIZAWA, K.; GILLETT, C. (2011). "The autonomy of psychology in the age of neuroscience". In: *Causality in the sciences*, P. M. Illari, F. Russo and J. Williamson (eds.) Oxford: Oxford University Press.

ARISTOTLE. (1963). *De Interpretatione*. J. L. Ackrill (transl.). Oxford: Oxford University Press.

ASHCRAFT, M. H. (1993). "A personal case history of transient anomia". *Brain and language*, 44, pp. 47-57.

AYDEDE, M.; (2015). "The Language of Thought Hypothesis", *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.).

BELLETTI, A.; RIZZI, L. (eds.). (2002). "Editors' introduction: some concepts and issues in linguistic theory". In: *On Nature and language.* N. Chomsky Cambridge: Cambridge University Press.

BERLIN, B.; KAY, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

BERMÚDEZ, J. L. (2003). *Thinking without words*. Oxford: Oxford University Press.

BERMÚDEZ, J. L. (2014). *Cognitive Science*. 2nd edition. Cambridge: Cambridge University Press.

BICKLE, J. (2013). "Multiple realizability", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.).

BLOCK, N. (1980). "What is functionalism?", In: Heil J. (ed.) *Philosophy of Mind: a guide and anthology*. Oxford: Oxford University Press, 2004.

BOHN, O-C. (2000). "Linguistic relativity in speech perception" In: *Evidence for linguistic relativity*. S. Niemeier and R. Dirven (eds). Amsterdam: John Benjamins Publishing Company.

BOONE, W.; PICCININI, G., (2016). "The cognitive neuroscience revolution". *Synthese* 193:1509–1534.

CAIN, M. J. (2002). *Fodor: Language, Mind and Philosophy*. Cambridge: Polity Press.

CALVO, P.; SYMONS, J. (2014). *The architecture of cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, MA: MIT Press.

CARRUTHERS, P. (1996). *Language, Thought and Consciousness: an essay in philosophical psychology*. Cambridge: Cambridge University Press.

CARRUTHERS, P. (2002). "The cognitive functions of language" in: *Behavioral and Brain Sciences*, 25:6, 657-674.

CARRUTHERS, P. (2006). *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

CARRUTHERS, P. (2009) "Invertebrate concepts confront the generality constraint (and win)". In: *The philosophy of animal minds*. Lurz, R. W. (ed.) Cambridge: Cambridge University Press.

CARRUTHERS, P.; BOUCHER, J. (eds) (1998). *Language and Thought: Interdisciplinary Themes*. Cambridge: Cambridge University Press.

CHENEY, D. L.; SEYFARTH, R. M. (1990). "The representation of social relations by monkeys". *Cognition*, 37, pp. 167-196.

CHOMSKY, N. (2009). *Cartesian linguistics*. 3rd edition. Cambridge: Cambridge University Press.

CLAPP, L. (2012). "Is even thought compositional?". *Philosophical Studies*, Vol. 157 Issue 2 pp. 299-322.

CLAPP, L.; DUHAU, L. (2011). "Varieties of the generality constraint". *Manuscrito*, v. 34, n. 2, pp. 397-433.

CRANE, T. (2003). *The mechanical mind: a philosophical introduction to minds, machines and mental representation* (2nd edition). Taylor & Francis e-Library.

CUMMINS, R. (1996). "Systematicity". In: *The world in the head*. Oxford: Oxford University Press, 2013.

CUMMINS, R. et al. (2001). "Systematicity and the cognition of structured domains". In: *The world in the head*. Oxford: Oxford University Press, 2013.

DEL PINAL, G. (2015). "The Structure of Semantic Competence: Compositionality as an Innate Constraint of the Faculty of Language". *Mind & Language*, Vol. 30, No. 4, pp. 375–413.

DESCARTES, R. (1637). *A discourse on the method*. I. Maclean (trans.) Oxford: Oxford University Press, 2006.

DESCARTES, R. (1649). "Letter to More". In: *The Philosophical writings of Descartes: The correspondence*, Vol III. Cambridge: Cambridge University Press, 1991.

DEVITT, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.

DUHAU, L. (2010). "Do bees really have concepts? A discussion of Carruthers' criteria for conceptuality". *Teorema: Revista Internacional de Filosofía* Vol. 29, No. 2, La mente de los animales: Animal Minds, pp. 125-134.

ELUGARDO, R., (2005). "Fodor's inexplicitness argument". In: M. Werning, E. Machery and G. Schurz (Eds.), *The compositionality of meaning and content. Volume 1: foundational issues*, Ontos Verlag, Berlin, pp. 59-85.

EVANS, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

EVANS, N. (2010). *Dying Words: Endangered Languages and What They Have to Tell Us*. Oxford: Wiley-Blackwell.

EVERETT, C. (2013). *Linguistic relativity*. Berlin: De Gruyter.

FODOR, J. A. (1974). "Special sciences". In: [Introduction of] *The Language of Thought*. Cambridge, MA: Harvard University Press, 1975.

FODOR, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

FODOR, J. A. (1980a). "Methodological solipsism considered as a research strategy in cognitive psychology". In: *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press Bradford Books, 1981.

FODOR, J. A. (1980b). "Searle on what only brains can do". In: *Behavioral and Brain Sciences* 3 (3):431.

FODOR, J. A. (1981a). "The Mind-Body Problem". In: *Scientific American*. 244:114-25.

FODOR, J. A. (1981b). "The present status of the innateness controversy". In: *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press Bradford Books.

FODOR, J. A. (1981c) *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press Bradford Books.

FODOR, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press Bradford Books.

FODOR, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press Bradford Books, 1993.

FODOR, J. A. (1989). "Making mind matter more". In: *Theory of content and other essays*. Cambridge, MA: MIT Press Bradford Books, 1992.

FODOR, J. A. (1990). *Theory of content and other essays*. Cambridge, MA: MIT Press Bradford Books, 1992.

FODOR, J. A. (1991a). "Afterthoughts: Yin and Yang in the Chinese room". In: *The nature of mind*. Rosenthal, D. M. (ed.) Oxford: Oxford University Press.

FODOR, J. A. (1991b). "Replies". In: *Meaning in Mind: Fodor and his critics,* B. Loewer; G. Rey (eds.). London: Blackwell, 1993.

FODOR, J. A. (1994). *The elm and the expert*. Cambridge, MA: MIT Press Bradford Books.

FODOR, J. A. (1997). "Special sciences: still autonomous after all these years". In: *In critical condition*. Cambridge, MA: MIT Press Bradford Books, 2000.

FODOR, J. A. (1998a). *Concepts*. Oxford: Oxford University Press.

FODOR, J. A. (1998b). "Do we think in Mentalese? Remarks on some arguments of Peter Carruthers". In: *In critical condition*. Cambridge, MA: MIT Press Bradford Books, 2000.

FODOR, J. A. (1998c). *In critical condition*. Cambridge, MA: MIT Press Bradford Books, 2000.

FODOR, J. A. (1999). "Diary". In: *London Review of Books*, Vol. 21 No. 19, pages 68-69.

FODOR, J. A. (2001a). "Language, thought and compositionality". In: *Mind & Language*, Vol. 16 No. 1, pp. 1-15.

FODOR, J. A. (2001b). *The mind doesn't work that way*. Cambridge, MA: MIT Press Bradford Books.

FODOR, J. A. (2003a). *Hume variations*. Oxford: Oxford University Press.

FODOR, J. A. (2003b). "Is it a bird?". In: *The times literary supplement*.

FODOR, J. A. (2003c). "More peanuts". In: *London Review of Books*,Vol. 25, No. 19, pp. 16-17.

FODOR, J. A. (2007). "Semantics: an interview with Jerry Fodor". ReVEL. Vol. 5, n. 8.

FODOR, J. A. (2008). *LOT 2*. Oxford: Oxford University Press.

FODOR, J. A.; BLOCK, N. (1972). "What psychological states are not". In: *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press Bradford Books, 1981.

FODOR, J. A.; LEPORE, E. (2002). *The Compositionality papers*. Oxford: Oxford University Press.

FODOR, J. A.; MCLAUGHLIN, B. (1990). "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work". In: *In Critical Condition*. Cambridge, MA: MIT Press Bradford Books, 2000.

FODOR, J. A.; PYLYSHYN, Z. W. (1988). "Connectionism and cognitive architecture: A critical analysis". *Cognition*, 28, 3–71.

FODOR, J. A.; PYLYSHYN, Z. W. (2015). *Minds without meaning*. Cambridge, MA: MIT Press.

FREGE, G. (1923). "Compound thoughts". *Mind*, New Series, Vol. 72, No. 285 (Jan., 1963), pp. 1-17

GARSON, J. (2015). "Connectionism", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.).

GILBERT, A. et al. (2006). "Whorf hypothesis is supported in the right visual field but not the left". In: *Proceedings of the National Academy of Sciences*, 103 (2) 489-494.

GRANDIN, T. (1995). *Thinking in Pictures: my life with autism*. 2nd edition. New York: Vintage Books.

GRICE, P. (1975). "Logic and conversation". In: P. Cole and J. L. Morgan (eds.) *Syntax and Semantics, Vol. 3, Speech Acts*. New York: Academic Press.

HEIDER, E. R. (1972) "Universals in color naming and memory". *Journal of Experimental Psychology*, Vol. 93, No. 1, 10-20.

HEIL, J. (ed.) (2004). *Philosophy of Mind: a guide and anthology*. Oxford: Oxford University Press.

HOLMES, K. J.; WOLFF, P. (2012). "Does Categorical Perception in the Left Hemisphere Depend on Language?". Journal of Experimental Psychology: General, Vol. 141, No. 3, 439–443.

HUTTO, D. D.; MYIN, E. (2013) *Radicalizing enactivism*. Cambridge, MA: MIT Press Bradford Books

KABADAYI, C.; OSVATH, M. (2017). "Ravens parallel great apes in flexible planning for tool-use and bartering". *Science,* Vol. 357, Issue 6347, pp. 202-204

KATZ, M. "The Language of Thought Hypothesis". In: *The Internet Encyclopedia of Philosophy*.

KAY, P.; KEMPTON, W. (1984). "What is the Sapir–Whorf Hypothesis?." *American Anthropologist* 86: 65–79.

KIM, J. (1992). "Multiple realization and the metaphysics of reduction". In: Philosophy and Phenomenological Research, Vol. 52, No. 1, pp. 1-26.

KIM, J. (1996). *Philosophy of mind*. Boulder: Westview Press, 1998.

KRIPKE, S. (1980) *Naming and necessity*. Oxford: Blackwell Publishing.

KRUPENYE ET AL. (2016). "Great apes anticipate that other individuals will act according to false beliefs". *Science*, Vol. 354, Issue 6308, pp. 110-114.

LI, P. et al. (2011). "Spatial reasoning in Tenejapan Mayans". *Cognition*, 120, pp. 33-53.

LIN, L. et al. (2007). "Neural encoding of the concept of nest in the mouse brain". *PNAS*, Vol. 104, no. 14, pp. 6066–6071.

MALT, B. C.; WOLFF, P. (eds.) (2010). *Words and the mind: how words capture human experience*. Oxford: Oxford University Press.

MARGOLIS, E.; LAURENCE, S. (2011). "Learning Matters: The Role of Learning in Concept Acquisition". In: Mind & Language, Vol. 26, No. 5, pp. 507–539.

OCKHAM, W. (1974). *Ockham's theory of terms: Part I of the* Summa Logicae. M. J. Loux (transl. and intro.) Notre Dame: University of Notre Dame Press.

PAGIN, P. (2012) "Communication and the complexity of semantics". In: M. Werning, W. Hinzen and E. Machery (eds.) *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press.

PANACCIO, C. (2017). *Mental language: from Plato to William of Ockham*. J. P. Hochschild and M. K. Ziebart (transl.). New York: Fordham University Press.

PEREBOOM, D.; KORNBLITH, H. (1991). "The metaphysics of irreducibility". In: *Philosophy of Mind: a guide and anthology*. J. Heil (ed.) Oxford: Oxford University Press.

PESSIN, A. (1995). "Mentalese syntax". In: *Philosophical Studies*, 78: 33-53.

PINKER, S. (1994). *The Language Instinct*. London: Penguin Books.

PINKER, S. (1997). *How the Mind Works*. New York: W. W. Norton & Company.

PINKER, S. (2005). "*So how* does *the mind work*?" In: *Mind & Language*, Vol. 20 No. 1, pp. 1–24.

PINKER, S. (2007). *The Stuff of Thought*. New York: Penguin Books.

PLACE, U. T. (1956). "Is consciousness a brain process?" In: British journal of psychology, 47:1, pp. 44-50.

PRINZ, J. (2012). "Regaining composure: a defence of prototype compositionality". In: M. Werning, W. Hinzen and E. Machery (eds.) *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press.

PUTNAM, H. (1963). "Brains and behavior", in: Heil, J. (ed.) *Philosophy of Mind: a guide and anthology*. Oxford: Oxford University Press, 2004.

PUTNAM, H. (1967). "Psychological Predicates", in: Heil, J. (ed.) *Philosophy of Mind: a guide and anthology*. Oxford: Oxford University Press, 2004.

PYLYSHYN, Z. (2003) *Seeing and Visualizing*. Cambridge, MA: MIT Press.

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

REGIER, T. et al. (2010). "Language and thought: which side are you on, anyway". In *Words and the mind: how words capture human experience*, B. C. Malt and P. Wolff (eds.). Oxford: Oxford

University Press.

ROBERSON, D. et al. (2000). "Color Categories Are Not Universal: Replications and New Evidence From a Stone-Age Culture". In: *Journal of Experimental Psychology*, Vol. 129, No. 3, 369-398.

ROBERSON, D. et al. (2005). "Color categories: Evidence for the cultural relativity hypothesis." *Cognitive Psychology*, 50, 378–411.

ROBERSON, D. et al. (2009). "Thresholds for color discrimination in English and Korean speakers". *Cognition*, 112, 482–487.

ROBERSON, D.; HANLEY, R. (2010). "Relatively speaking: an account of the relationship between language and thought in the color domain", in *Words and the mind: how words capture human experience*, B. C. Malt and P. Wolff (eds.). Oxford: Oxford University Press.

RYLE, G. (1949) *The concept of mind*. London: Penguin Books, 1990.

SEARLE, J. R. (1980a). "from Author's response". In: In: *The nature of mind*. Rosenthal, D. M. (ed.) Oxford: Oxford University Press, 1991.

SEARLE, J. R. (1980b). "Minds, brains and programs". In: Behavioral and Brain Sciences 3 (3): 417-457.

SEARLE, J. R. (1982). "The myth of the computer". In: *The New York review of books*.

SEARLE, J. R. (1983). *Intentionality*. Cambridge: Cambridge University Press, 1999.

SEARLE, J. R. (1990a). "Is the brain a digital computer?". Proceedings and Addresses of the American Philosophical Association, Vol. 64, No. 3, pp. 21-37.

SEARLE, J. R. (1990b). "Is the brain's mind a computer program?". In: *Scientific American*, 262 (1):26-31.

SEARLE, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press Bradford Books.

SHARKEY, A.; SHARKEY, N. (2009). "Connectionism", in: *The Routledge Companion to Philosophy of Psychology,* J. Symons and P. Calvo (eds.). New York: Routledge.

SKELTON, A. E. et al. (2017). "Biological origins of color categorization". PNAS. 114 (21) 5545-5550.

SMART, J. J. C. (1959). "Sensations and brain processes". In: The Philosophical Review, Vol. 68, No. 2, pp. 141-156.

SMOLENSKY, P. (1989). "Connectionist Modeling: Neural Computation/Mental Connections" in: *Mind Design II*. Cambridge, MA: MIT Press Bradford Books, 1997.

STRAWSON, G. (2016). "Consciousness isn't a mystery. It's matter." In: *The New York Times*.

SYMONS, J.; CALVO, P. (eds.) (2009). *The Routledge Companion to Philosophy of Psychology*. New York: Routledge.

SZABÓ, Z. G. (2000). "Compositionality as supervenience". *Linguistics and Philosophy*. Vol. 23, No. 5, pp. 475-505.

SZABÓ, Z. G. (2010). "The determination of content". *Philosophical Studies*. Vol. 148, pp. 253–272.

SZABÓ, Z. G. (2012). "The case for compositionality". In: M. Werning, W. Hinzen and E. Machery (eds.) *The Oxford Handbook of Compositionality.* Oxford: Oxford University Press.

TOMASELLO, M.; HERRMANN, E. (2010) "Ape and Human Cognition: What's the Difference?" *Current Directions in Psychological Science*, 19(1) 3-8.

TURING, A. (1950). "Computing machinery and intelligence". In: *Mind*, New Series, Vol. 59, No. 236, pp. 433-460.

WAGNER, K. et al. (2013) "Slow mapping: color word learning as a gradual inductive process". *Cognition*;127(3):307-17.

WASKAN, J. "Connectionism". In: *The Internet Encyclopedia of Philosophy*.

WHORF, B. L. (1940). "Science and linguistics" in *Language, Thought, and Reality: Selected writings of Benjamin Lee Whorf*. J. B. Carroll (Ed.). Cambridge: The MIT Press, 1956.

WINAWER, J. et al. (2007). "Russian blues reveal effects of language on color discrimination." In: Proc Natl Acad Sci. 2007; 104: 7780–7785.

WITTGENSTEIN, L. (1953). *Philosophical investigations*. Chichester: Wiley-Blackwell, 2009.

WITTGENSTEIN, L. (1958). *The blue book*. New York: Harper Torchbooks, 1965.

WOLFF, P.; HOLMES, K. J. (2011) "Linguistic relativity". In: *WIREs Cognitive Science*, Vol 2, Issue 3, pp. 253-265.

YANG, Y. et al. (2017). "Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function". In: *NeuroImage* 146, 658–666.